DOCUMENT RESUME

ED 279 730 TM 870 180

AUTHOR	Brandt, David A.
TITLE	Comparison of Computer Programs Which Compute Sampling Errors for Complex Samples. Technical Report 26.
INSTITUTION	American Institutes for Research in the Behavioral Sciences. Palo Alto, CA. Statistical Analysis Group in Education.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	AIR-87600-6/82-TR
PUB DATE	Jun 82
CONTRACT	300-78-0150
NOTE	74p.; Final page of Appendix 2 has faint blurred type and may not reproduce well.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Analysis of Variance; Cluster Analysis; *Computer Software Reviews; Correlation; Costs; *Error of Measurement; *Evaluation Criteria; Measurement Techniques; Regression (Statistics); *Sampling; *Statistical Analysis; Statistical Bias
IDENTIFIERS	Michigan Terminal System; *OSIRIS IV (Computer Program); *SUPER CARP (Computer Program)

ABSTRACT

This report describes and evaluates the major computer software packages capable of computing standard errors for statistics estimated from complex samples. It first describes the problem and the proposed solutions. The two major programs presently available, SUPER CARP and OSIRIS, are described in general terms. The kinds of statistics available from each program and the methods of solution each employs are discussed. The program documentation and ease of deck setup are compared, and the cost of acquiring and running the programs is presented. Appendix 1 contains technical papers on the programs and Appendix 2 describes the major features and options of one program more fully. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279730

Technical Report No. 26

Comparison of Computer Programs Which Compute Sampling Errors for Complex Samples

David A. Brandt

Prepared by

SAGE
STATISTICAL ANALYSIS GROUP IN EDUCATION

For the

National Center for Education Statistics

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. V. Clemans

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

American Institutes for Research

Box 1113, Palo Alto, California 94302

TM 810 180

BEST COPY AVAILABLE

AIR 87600-6/82-TR

TECHNICAL REPORT 26

COMPARISON OF COMPUTER PROGRAMS WHICH COMPUTE SAMPLING ERRORS FOR COMPLEX SAMPLES

Submitted to the National Center for Education Statistics

By

David A. Brandt

Statistical Analysis Group in Education
American Institutes for Research
P. O. Box 1113
Palo Alto, California 94302

This work was done under Contract No. 300-78-0150 with the National Center for Education Statistics, Department of Health, Education, and Welfare. However, the content does not necessarily reflect the position or policy of either angency, and no official endorsement should be inffered.

June 1982

ABSTRACT

This report describes and evaluates the major computer software packages capable of computing standard errors for statistics estimated from complex samples. We first describe the problem and the proposed solutions. The presently available software is then described in general terms. We then present in detail the kinds of statistics available from each program and discuss the methods of solution each employs. We then compare the program documentation and ease of deck setup. The cost of acquiring and running the programs is also discussed briefly. Appendix 1 contains technical papers on the programs and Appèndix 2 describes the major features and options of one program more fully.

TABLE OF CONTENTS

	Page
Introduction	1
Major Computer Programs for Computing Sampling Errors	3
Major Features and Options of OSIRIS IV and SUPER CARP	5
Methods of Computing Sampling Errors for Complex Samples	8
Taylor Series Expansion	8
Repeated Replications	9
Comparison of Accuracy of the Methods	11
Cost	12
Documentation and Ease of Use	13
Availability	23
Possible Modifications	23
Summary	24
References	27
Appendix 1. Technical papers on methods of solution in SUPER CARP and OSIRIS	29
Appendix 2. Description of OSIRIS program package	51

- - - - - - - 0 - - - - - - -

Table 1. Coverage of OSIRIS IV and SUPER CARP	6

Introduction

Methods of hypothesis testing based on normal statistical theory provide valid results when subjects are independently sampled from a well-defined population. However, the cluster sampling methods routinely employed in surveys do not, in general, provide independent observations. This is because the probability of inclusion of a particular case is greater given the inclusion of another unit from the same cluster. The inappropriate use of statistical methods for independent observations will lead to an underestimate of the actual variance. Thus, confidence bounds for statistics will be underestimated, and hypotheses may be falsely rejected by statistical tests based on such calculations. Conversely, the use of stratification variables can reduce the sampling variance for the variables correlated with the stratification variables. In such cases, the use of statistics derived for simple random samples will overestimate the error variance and create the opposite kinds of errors.

Kish (1965) has developed a statistic which expresses the practical effect of complex sampling. He defined the Design Effect (DEFF) as the ratio of the sampling error of an estimator taking into account complex sampling to the sampling error of the estimator assuming independent observations. Statistics computed from sample surveys typically have DEFF's greater than one (Kish & Frankel, 1974). There is a strong consensus among survey researchers that the problem is important and needs to be given serious consideration (Frankel, 1971; Fuller, 1975; Kalton, 1977).

Based on the large-scale empirical investigation of the effects of cluster sampling done by Frankel (1971; see also Kish & Frankel, 1974), we can make some general statements concerning its effect on sampling errors. In general, design effects are larger for simple statistics such as totals and ratios than for correlations and regression coefficients. Design effects of two or greater are not uncommon. Although design effects vary within a survey, the design effects for a given type of statistic (e.g., totals) are reasonably consistent within a survey. However, the typical design effect for totals may differ appreciably from the typical design effect for regression coefficients.

Use of standard errors computed assuming a simple random sample (SRS) results in a high Type I error rate; the methods for estimating standard errors from clustered samples disussed in this report, in general, are slightly conservative, but the magnitude of the error is small.

We know of four programs capable of computing standard errors for statistics computed from complex samples. Two of the programs, SUPER CARP and OSIRIS IV, are highly evolved and contain facilities for handling a wide range of problems. The two remaining programs are less well developed and therefore handle a more limited range of problems. It appears that any improvements to these programs would simply result in the duplication of features and options presently available in either or both SUPER CARP and OSIRIS IV. Thus, it seems that the choice should be between the two most completely developed programs. Before discussing SUPER CARP and OSIRIS IV in detail, we will mention the two other programs and indicate their (relative) shortcomings.

A program SURREG, written by Shah, et al., uses the Taylor series method (see below) to compute standard errors for regression coefficients. The disadvantage of SURREG is that it does not handle simpler statistics such as totals, means, ratios, proportions, and differences betweeen ratios (e.g., between subpopulations). In order to cover these areas, SURREG would have to be used in combination with either SUPER CARP or OSIRIS IV. But since SUPER CARP and OSIRIS IV already cover all the cases handled by SURREG, it would seem to be superfluous. Furthermore, SUPER CARP handles the same cases using the same method of solution and its coverage of regression is more complete.

We are also aware of a program developed at WESTAT for computing sampling errors for totals and ratios. It is written as a SAS procedure NASSVAR and uses balanced repeated replication methods (see below). We understand that the program was written to analyze a particular survey, but it is sufficiently flexible to be of some general interest. At present there is no manual, but appropriate documentation could be assembled. We were unable to test the program and, of course, could not review the documentation since no "official" documentation exists.

Our understanding is that the program is capable of computing a wide variety of simple estimates and their sampling errors, such as totals, means, ratios, differences between ratios. In its present form, however, it is much more cumbersome to use because the user must supply a design matrix which defines the replicates. PROC MATRIX must be called before NASSVAR can do its calculations. Furthermore, the balanced repeated replication technique requires that the sample be designed with two primary selections per stratum. Since no other methods are programmed, this feature would require the user to adjust the design of some surveys so that it conforms to this pattern. That is, regardless of the "actual" design, the data input to the program must have 2 PS per stratum, and an appropriate design matrix must exist for the number of strata in the survey.

In addition, in its current version, it cannot be used for doing more complicated analyses such as multiple linear regression. This is because the statistic to be estimated must be programmed by the user. Thus, it is really designed to handle univariate statistics only. However, if work on this program continues, it could reach the stage of development of SUPER CARP or OSIRIS.

Major Computer Programs for Computing Sampling Errors

SUPER CARP and OSIRIS IV are the only two programs that compute sampling errors for both simple and complex statistics. Both programs have been released for general use and one, OSIRIS, is a fully supported program package. Of the two, SUPER CARP is the more versatile and powerful. It has more features and options than OSIRIS IV and also differs in that it uses only one method for computing the sampling errors for all statistics. The more unified treatment of the problem in SUPER CARP means that it can handle surveys of any size and design within a single framework. The coverage of the OSIRIS IV program is also broad, particularly with regard to the computation of standard errors for simple statistics, but for regression problems the user has to attend to the method of solution appropriate for that problem. Its single most attractive method for regression problems, Balanced

Repeated Replications, is not available for "unbalanced" designs or for very large balanced designs.

OSIRIS IV is a major statistical program package developed specifically for the analysis of sample surveys. It features a free format control language, general and flexible data cleaning and file manipulation facilities, and numerous data analysis subprograms. The program contains facilities for storing variable self-descriptors which name and label the variables, define missing values, and name the values of the variable levels. This set of descriptors is known as the OSIRIS "dictionary" and is stored separately from the dataset. To carry out statistical analyses, both files are accessed. The package also contains SAS to OSIRIS and SPSS to OSIRIS interfaces, &SASFILE and &SPSSFILE. They permit the user to bring data stored in other systems into OSIRIS IV without having to create the self-described dataset manually. If several analyses in OSIRIS are anticipated, &SASFILE or &SPSSFILE may be used to create a permanent OSIRIS dataset by simply providing the appropriate I/O assignments. The program will then create and store the OSIRIS dictionary for later use. Furthermore, both SAS and SPSS can read OSIRIS datasets through PROC CONVERT and OSIRIS VARS, respectively.

OSIRIS IV is unique among the widely available packages in that the variable self-descriptors can be augmented with codebook records and stored online. The codebook may be printed out separately (e.g., for publication) or the portion of the codebook describing the variables used in a statistical analysis can be listed on the printout along with the results. This option is requested by simply setting the keyword PRINT equal to CODEBOOK on the parameter card.

A second unique feature of OSIRIS IV is that it contains highly evolved and well integrated facilities for reading, storing, modifying, and analyzing hierarchical datasets with variable length records. These facilities make it possible to bypass complex and difficult data management steps which have to be carried out manually on other packages. The OSIRIS IV package is described in greater detail in Appendix 2.

SUPER CARP is a stand-alone program written in FORTRAN G and is based on the method of solution given in Fuller (1975). The control language is fixed format, but since the program is more powerful and versatile than any

other, it is capable of carrying out analyses that would otherwise be impossible. These special features will be described in the following section. SUPER CARP is more difficult to use, both because the control language is fixed format and because it does not contain interfaces to any package programs.

The two programs, OSIRIS IV and SUPER CARP, use the same method of solution for simple statistics (Taylor series expansion) but differ in how they handle regression and correlation problems. SUPER CARP invariantly uses Taylor series, while OSIRIS IV uses repeated replication techniques. In OSIRIS IV, the user has control over the type of repeated replication tech- nique he or she wishes to use. Balanced repeated replications is a built-in and preferred option.

In OSIRIS IV the major features of SUPER CARP are split between two subprograms, &PSALMS and &REPERR. The methods of solution are described in Vinter (1980). &REPERR computes sampling errors for means, correlations, and regression coefficients, while &PSALMS handles totals, ratios, and differences between ratios. &PSALMS invariantly employs TAYLOR; &REPERR has built-in options for Balanced Repeated Replications (BRR), Jackknife Repeated Replications (JRR), and JRR with random selection of two primary selections per stratum. In addition, the user may define the replicates manually. This requires the preparation of k cards, each defining a replicate.

Major Features and Options of OSIRIS IV and SUPER CARP

Table 1 summarizes the statistics computed by these two programs; it is clear that there is considerable overlap in coverage between OSIRIS IV and SUPER CARP. However, SUPER CARP includes more complete facilities for regression problems. The option for regression with a nested error structure is appropriate for so-called "pooled cross-section" designs. They are common in econometric research and have also been used by sociologists such as Michael Hannon. The two-stage sample option computes the sampling variance of all requested statistics as the sum of two components.

Table 1

Coverage of OSIRIS IV and SUPER CARP

Statistic	SUPER CARP	OSIRIS IV &PSALMS	OSIRIS IV &REPERR
Totals	X	X	
Ratios	X	X	
Means	X	X	X
Difference between two ratios	X	X	
Subpopulation totals	X	X	
Subpopulation ratios	X	X	
Subpopulation Means	X	X	
Subpopulation proportions	X	X	
Test of independence (contingency table)	X		
Regression equations	X		X
Correlations	X		X
Standard errors of above statistics	X	X	X
Covariance matrix of parameter estimates	X		X
Design effect	X	X	
Standard errors assuming simple random sample	X	X	
Coefficient of variation		X	
Bias		X	
Covariance of numerator and denomenator		X	
Intraclass correlation		X	
Regression with nested error structure	X		
2-stage samples	X		
Errors-in-variables regression models	X		

Most notably, SUPER CARP contains several so-called "errors-in-variables" models based on the theoretical work of Fuller (1980; Fuller & Hidiroglou, 1978). These models assume that the predictor variables are measured with error; multiple linear regression ordinarily assumes that the predictors are measured perfectly. To use this feature, either the covariance matrix of the errors or the reliabilities of the predictor variables must be known (e.g., from previous research). The user must input the matrix or a function of the reliabilities. SUPER CARP does not estimate the reliabilities of the predictors, as does a program such as LISREL. It does allow the user to provide estimated, rather than known, error variances for the predictors. In effect, the program disattenuates the regression equation.

In general, the effect of measurement error is a reduction of the absolute level of the regression coefficient in relation to its expected value in the absence of error (Fuller, 1975). This feature of SUPER CARP is also of obvious value in studies using simple random sampling; it is possible to use the errors-in-variables option in the program independently of the options for handling clustered sampling. Thus, SUPER CARP handles two problems which frequently arise in survey work: cluster sampling and errors-in-variables.

OSIRIS IV, on the other hand, computes several useful diagnostic statistics that are missing from SUPER CARP's output. These include the Design Effect, the square root of the Design Effect, the bias of a ratio estimate, the coefficient of variation (if the coefficient of variation of the denomenator is greater than .15 a warning is printed), and the intraclass correlation. The Design Effect, in particular, is a well-known statistic that is frequently reported as an indicator of the effect of the cluster sampling. To obtain the Design Effect using SUPER CARP, one has to rerun the problem in a standard package program and form the ratio of standard errors manually from the information in both printouts.

Taylor Series Expansion

The first method of estimating sampling errors is the Taylor series expansion of the first-order statistic (TAYLOR). This method relies on a model and makes certain assumptions regarding the adequacy of the model. Using this method, a statistic is expressed as a polynomial using the Taylor series expansion. The approximate variance of the statistic can then be obtained by using only the linear terms of this expansion.* In computing an estimate of the variance of the mean, for example, this approach recognizes that there is variance associated with both the summed variable as well as the count variable. The formulas provide a method of approximating this variance estimate. Appendix 1 contains the paper by Professor Fuller which gives the expressions used in SUPER CARP.

In the past the major difficulty with TAYLOR was that it was difficult to apply for more complicated statistics such as correlations and regression coefficients. The general formulas are given in Tepping (1968), but these expressions are cumbersome for complex statistics. This has limited the use of TAYLOR to simple statistics where it performs effectively. We will refer to the implementation of TAYLOR using the analytic formulas of Tepping as TAYLOR-A.

---

* The Taylor series expansion is a classical method for approximating a function and is used extensively in numerical analysis. The basic idea is that a function, f, of a variable (e.g., f can be the mean) can be approximated by the polynomial series

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \ldots \frac{f^{(n-1)}(0)}{(n-1)!}x^{n-1} + E$$

Where $f(0)$ is the value of the function at zero
$f'(0)x$ is the first derivative of x at zero, etc. and
E is the remainder.

Extensive work with the Taylor series method has indicated that a satisfactory approximation to f can frequently be obtained by using only the linear terms of the series; that is, terms in the polynomial beyond the first derivative are ignored. The sampling variance of a first-order statistic, such as a total, is estimated from the sampling variance of the first degree terms of the Taylor series expansion. The main assumption of the model is that this constitutes an adequate approximation.

The OSIRIS IV subprogram &PSALMS uses TAYLOR-A to estimate sampling errors of ratios, ratio means, totals, and differences between ratios (e.g., differences between subclasses). The ratios are typically means or proportions. A second OSIRIS IV subprogram, &REPERR, uses repeated replication methods to handle means, correlations, and regression coefficients.

Recently, numerical methods of approximating the covariance matrix of regression coefficients have been developed (Fuller, 1975; Woodruff & Causey, 1976). This work makes it considerably easier to implement TAYLOR on the computer. Prior to this, TAYLOR had been regarded as a reasonable method only for simple linear estimators such as means and totals, since the analytic expressions of Tepping posed considerable computing difficulties. But the numerical methods make TAYLOR feasible for other statistics. The program SUPER CARP makes use of the simple numerical estimates given by Fuller (1975). We will refer to this method as TAYLOR-N.

Repeated Replications

The second approach is known as "repeated replications." Unlike the Taylor series expansion, this method is not model dependent and relies on the "brute force" computing power available on modern computers. With this approach, the variance of a statistic is estimated using the variability among k replicates of the full study. Each replicate is created by excluding a subsample of primary selections (PS) in the dataset. The idea is for each replication to reproduce, except for size, the design of the entire study. The statistic of interest is then estimated for the whole sample and for each replication. The variability among these estimates is used to estimate the variance of the statistic. Naturally, the precision of this estimate increases as more and more replicates are created. For large designs with many strata and PS's, the cost of using all possible replicates can become excessive. For that reason, several ways of selecting certain replicates have been developed. The aim of these strategies is to obtain as precise an estimate of the variance as possible with the fewest replicates.

The best known and most attractive method of selecting replicates is known as balanced half-samples or Balanced Repeated Replications (Kish & Frankel, 1970). BRR is actually a strategy for both survey design and analysis. It requires exactly two PS per stratum. Each replicate (half-sample) is formed by taking one of the two PS's in each of the H strata.

The replicates are selected in such a way that they are mutually orthogonal. This pattern of selection is based on sets of orthogonal design matrices (Hadamard matrices) developed by Plackett and Burman (1946). By choosing balanced half-samples, the number of replicates needed to gain full precision can be reduced considerably. With two PS in each of H strata, a total of 2^H possible replicates exist. However, the maximum number of balanced half-samples needed is the smallest number equal to or greater than H which is divisible by four. McCarthy (1966) showed formally that, for linear estimators, such balanced replications give the same precision as can be obtained by taking all 2^H replications. This result does not hold exactly for more complicated estimators (e.g., correlations, regression coefficients, etc.) but Kish and Frankel (1974) conjecture that it holds approximately.

The &REPERR program uses the Hadamard matrices in Plackett and Burman. This restricts the use of BRR to datasets with between 4 and 88 strata. This was the largest Hadamard matrix known in 1946. Since then other methods of constructing Hadamard matrices have been developed so that BRR, in theory, can be extended to other designs. However, these more recently discovered matrices are not built into &REPERR. Currently the only way to use them is to read the design matrix in manually. Obviously the program could be modified to incorporate these matrices. The following discussion describes some of these new applications.

Because of the attractive properties of orthogonal balance, several workers have extended this idea to other designs. Gurney and Jewett (1975) have generalized BRR to the case of p PS per stratum, where p is any prime number. They give a numerical example that employs 110 strata and three PS per stratum. We will refer to this method as Generalized BRR (GBRR).

Two workers have addressed the problem of reducing the number of required replications when H is large. Mellor (1973; see also Cochran, 1979) and Lee (1972) have developed a class of so-called Partially Balanced Repeated Replications (PBRR). PBRR can be applied with two PS per stratum, or, when combined with the methods of Gurney and Jewett, p PS per stratum. First, the total number of strata is divided into groups with approximately H/g strata per group. Then a fully balanced set of replications is used within each group. The resulting complete set of replications is not fully balanced (across groups) but is balanced within groups. Precision is highest when g is chosen small (e.g., g=2), but costs increase as g decreases.

Lee (1973) gives some guidelines on the selection of the PBRR designs in comparison to the corresponding BRR design.

Also, Hadamard matrices larger than order 88 have been discovered. Currently, Hadamard matrices of order 200+ have been shown to exist, although there are gaps in the sequence. Thus, the conventional design of two PS per stratum can be analyzed by BRR when the number of strata is larger than 88.

When it is impossible to apply any of the balanced or partially balanced methods because the design itself is unbalanced, the Jackknifing method is available (JRR). With this method, replicates are formed by omitting one PS per replication and reweighting the other PS's in that stratum, until all PS have been omitted. Thus, it can be applied when there are different numbers of PS within the strata. However, the resulting replicates are not orthogonal. &REPERR contains built-in options for JRR and for JRR with random selection of two PS per stratum, an option intended for use with designs with a large number of strata. Using this option, &REPERR can handle designs with many more than 88 strata.

Comparison of Accuracy of the Methods

A large scale study of TAYLOR-A, BRR, and JRR was undertaken by Frankel (1971; see also Kish & Frankel, 1974). Frankel used an actual dataset (the Current Population Survey) and evaluated the methods on both simple linear statistics as well as regression coefficients and correlations. He also varied the number of strata in the design (6, 12, and 30). The performance of these methods was evaluated using total error and distribution of t-ratios as criteria. All three methods gave good results. TAYLOR-A gave slightly lower mean squared error but slightly higher bias, while BRR had the most accurate distribution of t-ratios. No method was clearly better or worse than the others.

Woodruff and Causey (1976) compared TAYLOR-N with the three methods studied by Frankel. They also used the CPS dataset studied by Frankel. However, they extended the investigation to include designs with many more strata. They reported results for 6, 12, 30, 90, 270, and 810 strata designs. These findings are consistent with Frankel's: BRR again was the best using the t-ratio criterion and Taylor series methods the worst. No

clear differences between TAYLOR-N and TAYLOR-A were found. TAYLOR-A was very slightly superior to TAYLOR-N for the cases in which both were computed. However, the performance of TAYLOR-N improved as the number of strata increased. Using the distribution of t-ratios criterion, the performance of TAYLOR-N was excellent, when the design had 90 or more strata. This is significant since repeated replication methods become costly for large designs.

The most important message from this work is that all the above methods provide good results. It is not reasonable to select or rule out any program solely on the basis of the accuracy of its estimation method. The research reviewed here shows that the differences among competing methods are slight and are dependent on features of the particular design as well as criterion used. TAYLOR series methods do worse than BRR when the design has few strata, but this difference disappears as the number of strata increases. Furthermore, TAYLOR-N methods are both available and economically attractive when the design is larger than can be analyzed by BRR. Also, the research to date has only used a single dataset; characteristics of the dataset such as the degree of interdependence among the observations (i.e., the intraclass correlation) can be expected to affect the performance of the estimators. Because these methods are designed to deal with a problem that does not lend itself to "airtight" analytic investigation, it is not really possible to make a definitive statement regarding the superiority of one of the methods.

It should also be mentioned that Lee (1973) reported on the performance of PBRR in comparison to BRR. He found that the "best" PBRR designs sacrificed little in precision (about 12%), but an improperly constructed PBRR could be extremely inefficient. No comparisons between replication techniques and TAYLOR were reported.

Cost

In repeated replication methods, both cost and accuracy is a function of the number of PS (i.e., the number of replicates). It is thought that somewhere between 40 and 100 replicates are needed to achieve acceptable precision. With large problems the cost can be high. The cost of the TAYLOR

method is relatively independent of the number of PS's. SUPER CARP, in particular, is quite inexpensive--certainly on par with the cost of a comparable analysis in a SPSS regression. In this regard, SUPER CARP appears to be superior to &REPERR.

We were unable to formally test all the programs, mainly because the OSIRIS package is presently not available to us. To do a proper evaluation of the costs of running each program it is necessary to install both on the same computer, preferably the computer NCES intends to use. To do this, the OSIRIS package must be acquired. However, we have set up a computer account on the Michigan computer system if some testing is called for prior to acquisition.

SAGE owns a copy of SUPER CARP. We tested it on a NAEP dataset consisting of approximately 1000 cases and 110 strata. A regression analysis was carried out on both SUPER CARP and SAS using a dozen predictors. The cost of the SUPER CARP job was comparable to the SAS run, despite the extra computations carried out by SUPER CARP. In this respect, its performance was most impressive since a problem with a large number of strata can be expensive when carried out using repeated replication techniques.

Documentation and Ease of Use

OSIRIS IV has been written to run interactively under MTS (Michigan Terminal System), the special operating system at the University of Michigan. Thus, considerable effort has already been expended in the area of user convenience. The package features a free format keyword oriented control language, many default options, and mnemonic devices. The language is similar to that of SAS with the following major differences: (1) the "&" is used instead of the term "PROC", (2) parameters appear on a separate card instead of on the first card, and (3) the ";" is not used as a delimiter. ("Nothing" is the default delimiter, but the user can set "something", including the ";", as the delimiter.)

User convenience has not been given any priority in the development of SUPER CARP. Its main virtue is that it contains "state of the art" features for analyzing complex survey data. Unfortunately, the deck setup is entirely fixed format and does not employ default options to handle standard cases.

The labeling of variables and program output is more limited in SUPER CARP. Variable names are limited to 8 characters and no value labels are permitted. The program title is limited to 16 characters. OSIRIS variable names can be up to 24 characters and value labels up to 8 characters. The program title can be up to 100 characters and an unlimited number of comment statements can be used. Furthermore, supplementing the OSIRIS dictionary with codebook records permits virtually unlimited documentation of the variables. There is no limit to the number of codebook records per variable, and provisions have been made for storing a great deal of information on each variable, such as frequency distributions and miscellaneous remarks concerning problems with variables (e.g., bad coding, inconsistent coding, etc).

The current documentation for the two programs reflects their differing origins and intended audience. The SUPER CARP manual is longer and more complete than the &REPERR and &PSALMS writeups and contains all the technical details on the methods of solution used in the program. It is nearly 200 pages long and contains several example problems. The examples include an explanation of the problem, the deck setup, and the output. Instructions for setting up the control cards are complete and fairly easy to follow, but since it is a fixed format program with many features and options it is easy to make errors (e.g., entering a number in the wrong column or inadvertently requesting an unwanted option). Checking a deck setup must be done with the manual at hand since no mnemonic devices are used. Clearly, the documentation is intended for the experienced programmer who is also knowledgeable in statistics and is not afraid of technical statistical jargon.

The manual is divided up into sections on program input (i.e., deck setup), program algorithms, miscellaneous uses (e.g., special applications such as contingency tables), and examples. The section on program input is similar in style to the writeup for the old BMD (fixed format) package. For example, the user selects the type of analysis by specifying a number, e.g., "1" for regression analysis, "3" for total estimation, etc.

OSIRIS IV is intended for a larger and less technically sophisticated audience. The documentation excludes the technical details on the methods of solution (i.e., formulas) but this information is available elsewhere. The style of the documentation is more similar to SAS than other packages, but it is of much higher quality. That is, essential details are not

omitted and the package is more logically thought out. Each subprogram writeup is structured so that it is easy to use the manual as a reference. The sections within each subprogram writeup are:

(1) General Description - A short paragraph describing the main function of the program.

(2) Special Terminology - a glossary for technical terms used in the writeup, e.g., "Design Effect".

(3) Command Features - describes in detail the features and options in the program.

(4) Special features (optional) - describes extra features of program such as writing residuals to an output file.

(5) Printed Output - describes every statistic appearing in the output with page estimates.

(6) Input data - describes the options for the input of data, e.g., raw data or matrix input.

(7) Restrictions (optional) - limits, if any, to size of problem.

(8) Control Statements - Gives the directions for setting up the control cards.

(9) Examples - example deck setups and explanation of the problem.

Several example problems are given for each program. The examples include a brief description of the analysis and the deck setup only (no output). However, the previous version of OSIRIS provided a separate volume of the manual that contained sample outputs only. Perhaps OSIRIS IV well eventually be so documented.

To further demonstrate the control language of these programs, we now illustrate how each program can be used to do the same calculations. This is, essentially, the first example in the SUPER CARP manual.

The following SUPER CARP deck setup illustrates the estimation of totals, ratios, and multiple regression equations. This is followed by an OSIRIS deck setup which computes the same statistics.

```
EXAMPLE 1                9 1        4       1 0  41
(FORTRAN FORMAT CARD)
Y        X1        X2       X3
 0
  .10       .05
        9
   (DATA CARDS)
      3   2
      1   2
      2   2
      1   5
      1   4   1   1   1
      1   5   3   4
      2
      3   4
      1   3   1   0
      1   2   4
  (ENDFILE)
```

The explanation of these control cards is given in the following four pages, taken from the manual. The control language is typical of specialized stand-alone statistical programs: A few cards are used to "set up" the problem (Parameter card, format card, variable label card) and these are followed by an unlimited number of "analysis packets" which request specific data analyses.

Explanation of Super Carp as Taken from the Program Manual

Record 1: PARAMETER CARD

Columns	Coding	Comments
1-16	Example No. 1	Title for output
20-24	9	Number of observations
26	1	Stratum sampling rates will be read in
34-35	4	Number of variables read in
42	1	Program will generate variable 5 as a column of 1's, variable 5 will be named INTERCPT
44	0	Full survey type - (stratum, cluster and weight read in)
46-47	4	Number of analyses to be performed
48	1	All data to be listed in output

Record 2: FORMAT CARD

Columns	Coding
1-21	(2(I1, 1X), 5(F2.0, 1X))

Record 2a: VARIABLE NAME CARD

Columns	Coding	Comments
1-8	Y	Dependent variable
9-16	X1	
17-24	X2	Independent variables
25-32	X3	

Record 3: SCREENING CARD

Column	Coding	Comments
2	0	No screening required

Record 4: SAMPLING RATES CARD 1

Columns	Coding	Comments
1-8	0.10	Sampling rate for stratum 1
9-16	0.05	Sampling rate for stratum 2

Record 4: SAMPLING RATES CARD 2

Column	Coding	Comments
8	9	Indicates end of sampling rates

DATA

Data are entered on Cards in the specified input format.

Record 5: ANALYSIS CARD 1

Columns	Coding	Comments
4	3	Total estimation
8	2	Number of variables to be included in analysis

Record 6: VARIABLE IDENTIFICATION CARD 1

Column	Coding	Comments
4	1	Y total to be obtained
8	2	X1 total to be obtained

Record 5: ANALYSIS CARD 2

Columns	Coding	Comments
4	2	Ratio estimation
8	2	Number of variables to be included in analysis

Record 6: VARIABLE IDENTIFICATION CARD 2

Columns	Coding	Comments
4	1	Y is variable in numerator of the ratio
8	5	INTERCPT is variable in denominator of the ratio

Record 5: ANALYSIS CARD 3

Columns	Coding	Comments
4	1	Regression
8	4	Number of variables to be included in analysis
12	1	Weighted least squares
16	1	Intercept indicator is required
28	1	1 F-test is required

Record 6: VARIABLE IDENTIFICATION CARD 3

Columns	Coding	Comments
4	1	Y is dependent variable
8	5	INTERCPT is always listed after dependent variable
12	3	X2
16	4	X3

Other independent variables

Record 6c: NUMBER OF COEFFICIENTS TO BE TESTED

Columns	Coding	Comments
4	2	2 coefficients are to be tested

Record 6d: IDENTIFICATION OF COEFFICIENTS TO BE TESTED

Columns	Coding		
4	3	X2	Variables whose regression
8	4	X3	coefficients are to be tested

Record 5: ANALYSIS CARD 4

Columns	Coding	Comments
4	1	Regression
8	3	Number of variables to be included in analysis
12	1	Weighted least squares
16	0	No intercept in the regression

Record 6: VARIABLE IDENTIFICATION CARD 1

Columns	Coding	Comments
4	1	Y is dependent variable
8	2	X1
12	4	X3

independent variables

The OSIRIS deck setup is more typical of a modern program package. A convention of the OSIRIS package is that variables are identified by variable number in the deck setup and by name in the output. The OSIRIS dictionary is included to clarify the meaning of the control cards. Comments on each control card are contained in brackets.

```
&DICT                                          [Calls the program which creates
                                               the self descriptors]
Create self-descriptors for the SUPER CARP example [Label card]
PRINT=DICT                                     [Parameter card - print input
                                               dictionary]
V=1 NAME=STRATUM COL=1                         [Variable descriptors
                                               indicate starting
V=2 NAME='PRIMARY SELECTION' COL=3             column number and variable
V=3 NAME='CASE WEIGHT' COL=5                   width. Width defaults to 1
V=4 NAME='DEPENDENT VARIABLE' COL=8 W=2        if omitted. COL spec defaults
V=5 NAME='PREDICTOR 1'                         to previous spec plus value
V=6 NAME='PREDICTOR 2'                          of WIDTH]
V=7 NAME='PREDICTOR 3'
&END              [End of &DICT specs. Optional in batch mode, needed
                  in interactive mode]
&RECODE                        [Invokes OSIRIS recode facility]
R1=1                           [Variable R1 is set equal to constant 1]
&END
&PSALMS                                        [Invokes &PSALMS]
ratio and total estimation                     [Title card]
WTVAR=V3 SECU=V2 STRATUM=V1                     [parameter card identifies
                                                first three vars]
STRATA=1,2 SECU=6,3 MODEL=MULT                  [Use strata 1 and 2 containing 6
                                                and 3 SECU's]
NAME='TOTAL Y' PAR=V4/R1 CONSTANT=9   [Each card defines a population
NAME='TOTAL PREDICTOR 1' PAR=V5/R1 CONSTANT=9     parameter to be estimated.
NAME='MEAN Y' PAR=V4/R1                          It is defined using the
                                                 PAR keyword]
&END
&REPERR                                            [Invokes &REPERR]
multiple linear regression using repeated replications
STRATA=V1 SECU=V2 WTVAR=V3 VARS=V4-V7 PRINT=REP -  [Parameter card.Dash
                                                   is used to indicate
                                                   continuation]
STATS=(MEANS,CORR,MULTR)
STRATUM=1,2 SECU=6,3 MODEL=JKM     [Defines strata and model to be used]
DEPV=V4 VARS=V5-V7                 [Specifies the vars. in regression]
```

Aside from the differences caused by the fixed format versus keyword based control cards, several other features of the programs merit comment. The underlying structure of SUPER CARP is such that it could be converted to

a very convenient program by replacing the fixed format control language. A single method is used to handle all types of problems, and it is easy to request a wide variety of different analyses for the whole sample as well as any number of subpopulations in the same run. However, the lack of a free format keyword oriented control language makes access to these features relatively inconvenient.

OSIRIS &PSALMS appears to have already capitalized on the power of the Taylor series method as well as the ease of use of a modern keyword language. This particular example hardly begins to illustrate the range of analyses that can be conveniently carried out with the program. As with SUPER CARP, a single run can calculate any number of different estimates for both the total sample and an unlimited number of subpopulations. In &PSALMS both the estimates and the subpopulations can be labeled, supplementing the labeling provided by the dictionary and codebook.

The &REPERR program permits the user to request an unlimited number of regression analyses within a single run. Each regression analysis may be given its own label (up to 32 characters). However, it does not permit the user to analyze different subsets of the data within the same run, as does &PSALMS and SUPER CARP. To accomplish this, the user must invoke the &REPERR program once for each subset of the data he or she wishes to analyze.

When the BRR method is not being used, &REPERR is slightly less convenient to set up because the user has to provide the number of primary selections within each stratum (The BRR option, of course, assumes there are two). For very large designs this can become tedious. SUPER CARP does not require the user to furnish this information. This, of course, is still far more convenient than a repeated replications program such as the WESTAT program, which, at present, requires the user to input the Plackett-Burman design matrix for every problem.

Both OSIRIS programs, of course, permit the user to analyze any subset of the total number of variables in the dataset. The variables denoting the strata, primary selections, and case weights can appear anywhere in the file and are identified on the parameter card. SUPER CARP, at present, requires these variables to appear first in the dataset and in that order; they are identified by their location in the file.

Availability

We already have the SUPER CARP program, manual, and relevant literature. It can easily be installed on any IBM or IBM compatabile machine such as COMNET or NIH. For another AIR project, we installed it on the Stanford IBM 370/3081 without difficulty.

OSIRIS IV is available from the Institute for Social Research. The cost of the program and terms of the agreement are given in Appendix 2. It also can be installed on an IBM or IBM compatable machine. OSIRIS IV operates in batch mode under a conventional operating system such as at NIH or COMNET.

Another option is to use OSIRIS IV at the University of Michigan via TELENET. However, this would entail moving from one system to another and create logistical problems in obtaining hardcopy output. The advantage is that OSIRIS IV runs in a true interactive environment at Michigan, and JCL system commands have been replaced with a simple mnemonic keyword language. No system commands at all are needed to access disk files, because the operating system is capable of reading I/O assignment statements on program control cards. Tape files may be read after issuing the $MOUNT command. When using the computer interactively, system (MTS) commands may be issued at any point in the job stream; control will automatically be returned to MTS and then be returned to the program package when the requested task is completed.

Possible Modifications

Of course, the major disadvantages of SUPER CARP are its antiquated control language and its status as a "stand-alone" program. These problems would be completely overcome if SUPER CARP were brought into SAS as a procedure. In the fall (1981) Professor Fuller told us that SAS is, indeed, planning to incorporate SUPER CARP into their program as a fully supported subprogram. However, they indicated that the conversion would not take place until they had hired an additional programmer. We attempted to contact Professor Fuller in June regarding any progress made by SAS. We were told that he was out of the country until July, but a graduate student involved in the development of SUPER CARP told us that he had interviewed

for the SAS position and it was still not filled. Thus, work on converting SUPER CARP to an SAS procedure apparently has not started.

Converting SUPER CARP to an SAS procedure is a task SAGE can undertake if NCES indicates a need for its additional features. However, we will want to discuss this with Professor Fuller upon his return to the U.S. before taking responsibility for beginning this work.

Summary

Clearly, the two major programs reviewed here appear to be very sound alternatives. Both programs compute sampling errors for a wide variety of statistics, use proven methods of solution, and are adaptable to a wide variety of sampling designs. However, each program's special strengths lie in different areas. SUPER CARP's use of the Taylor series expansion means that it can handle surveys of any size and design without the user's having to worry about choosing an appropriate method or provide the program with additional information about the survey design. TAYLOR-N seems to be an especially sound method for doing multiple linear regression on large designs, e.g., over 100 strata. If NCES has a special need for regression analyses on these large designs, SUPER CARP would appear to be the program of choice. Of course, SUPER CARP also contains facilities for doing specialized analyses such as errors-in-variables models that are not available in any other program.

OSIRIS IV, on the other hand, is far stronger in the areas of user convenience and clarity of the documentation (to the non-sophisticated user). The &SASFILE and &SPSSFILE commands give users immediate and easy access to data stored in other systems and data stored initially in OSIRIS can easily be analyzed in SAS or SPSS via their interfaces. The keyword oriented control language makes the program much easier to learn and use. &PSALMS, in particular, is a very attractive program, because it combines the power and flexibility of the Taylor series method with the keyword oriented control language of a modern package program. If NCES's needs lie mainly in the areas covered by &PSALMS, it would seem to be the program of choice. &REPERR can be relatively tedious to set up for large unbalanced problems, but the program is capable of handling a wide variety of designs. The built in

options for Jackknife repeated replications and Jackknife repeated replications with random selection of two PS per stratum provide for regression analyses of large unbalanced designs; however these analyses are likely to be more expensive than the corresponding analyses in SUPER CARP.

Because OSIRIS is designed specifically to manage and analyze large-scale surveys, it contains other special features that are especially valuable to the survey researcher. The package is very obviously more fully self-documenting than any other program. The dictionary and codebook records may be stored jointly and accessed by any analysis program so that each run is fully and permanently documented. Also, the package offers very efficient ways of storing and managing large files. The portions of the program dealing with so-called structured files permit the user to store hierarchical datasets by storing each datapoint only once and without padding a dataset with missing values to make it rectangular. Furthermore, the package offers a wide variety of storage modes to maximize efficiency. For example, questionnaire response data can be stored in half-byte integer mode (4 bits); the most compact storage mode for numeric data available in SAS is two-byte (16 bits). For a dataset consisting of hundreds of such variables, the reduction in storage costs can be considerable.

Many of these advantages of OSIRIS would be offset if SUPER CARP were brought into SAS. In the long run, it appears very likely that this will happen. In the short run, however, it is unlikely that the staff of SAS will complete the conversion. It is in this area that SAGE may be able to help if NCES decides that SUPER CARP best meets their needs.

It is emphasized that since no method of solution has been shown to be clearly more accurate than another, this should not be a major deciding factor. In fact, program authors tend to choose one method over another for reasons of convenience rather than quality. It was decided to use Taylor series methods in SUPER CARP because of its great flexibility. It can be applied to any size survey no matter how large or unbalanced. OSIRIS uses Taylor series methods in &PSALMS for largely the same reasons, but used repeated replications to handle regression problems because those workers did not have access to the numerical methods developed by Fuller or Woodruff and Causey. Finally, repeated replication methods were used by WESTAT because it is a simple "brute force" procedure that can be applied to any statistic the user may care to estimate. Our judgment is that a choice of

software should be made on the basis of NCES's specific needs (e.g., types of statistics needed, design of surveys, sophistication of programmers, etc.) rather than on the basis of any "intrinsic" superiority of one approach over another.

As a final note, it should be recognized that our testing of the programs has been necessarily limited due to the short time available to prepare this report. Prior to making a final selection of programs, some Monte Carlo simulations to evaluate the accuracy of the methods or analyses of real data could be conducted using all available programs. Such testing would help determine the costs of using each program as well as the value of various unique features available in each package.

References

Brillinger, D.R. The application of the jackknife to the analysis of sample surveys. Commentary, 1966, 8, 74-80.

Casady, R.J. The estimation of variance components using balanced repeated replications. Proceedings of the Social Statistics Section, American Statistics Association, 1975, 352-357.

Cochran, W. Sampling techinques (3rd ed.). New York: Wiley, 1977.

Frankel, M. Inference from survey samples? Ann Arbor: Institute for Social Research, 1971.

Fuller, W.A. Regression analysis for sample survey. Sankhya C., 1975, 37, 117-132.

Fuller, W., & Hidiroglou, M. Regression estimation after correcting for attenuation. Journal of the American Statistical Association, 1978, 73, 99-104.

Gurney, M., & Jewett, R.S. Constructing orthogonal replications for variance estimation. Journal of the American Statistical Association, 1975, 70, 819-821.

Hidiroglou, M., Fuller, W., & Hickman, R. SUPERCARP. Ames, IA: Statistical Laboratory, 1980.

Kalton, G. Practical methods for estimating survey sampling errors. Bulletin of the International Statistical Institute, 1977, 43(3), 495-514.

Kish, L. Survey sampling. New York: Wiley, 1965.

Kish, L., & Frankel, M. Inference from complex samples. Journal of the Royal Statistical Society, 1974, 86, 1-37.

Kish, L., & Frankel, M. Balanced repeated replications for standard errors. Journal of the American Statistical Association, 1970, 65, 1071-1094.

Kish, L., Frankel, M., & Van Eck, N. SEPP: Sampling Error Program Package. Ann Arbor: ISR, 1972.

Kish, L., Nanboodri, K., & Pillai, R.K. The ratio bias in surveys. Journal of the American Statistical Association, 1962, 57, 863-876.

Koch, G., Freeman, D.H., & Freeman, J.L. Strategies in the multivariate analysis of data from complex surveys. International Statistical Review, 1975, 43(1), 59-78.

Koch, G., & Lemeshow, S. An application of multivariate analysis to complex sample survey data. Journal of the American Statistical Association, 1972, 67, 780-782.

Lee, K.H. Partially balanced designs for half-sample replication method of variance estimation. Journal of the American Statistical Association, 1972, 67, 324-334.

McCarthy, P.J. Replication: An approach to the analysis of data from complex surveys (Vital and Health Statistics, Series 2, No. 14). Washington, D.C.: HEW, 1966.

Mellor, R.W. Subsample replication variance estimation. Unpublished Dissertation, Harvard University, 1973.

Nathan, G. Tests of independence in contingency tables from stratified proportional samples. Sankhya,C. 1975, 37, 77-87.

OSIRIS. OSIRIS IV: Statistical Analysis and Data Management Software System. Ann Arbor: ISR, 1981.

Plackett, R.L., & Burman, P.J. The design of optimum multifactorial experiments. Biometrika, 1946, 33, 305-325.

Tepping, B.J. The estimation of variance in complex surveys. Proceedings of the Social Statistics Section, American Statistics Association, 1968, 11-18.

Vinter, S. Survey sampling with OSIRIS IV. Unpublished paper, 1980.

Woodruff, R.S., & Causey, B.D. Computerized method for approximating the variance of a complicated estimate. Journal of the American Statistical Association, 1976, 71, 315-321.

Appendix 1

Technical papers on methods of solution in SUPER CARP and OSIRIS

Sankhyā : The Indian Journal of Statistics
1975, Volume 37, Series C, Pt. 3, pp. 117-132.

REGRESSION ANALYSIS FOR SAMPLE SURVEY

By WAYNE A. FULLER*
Iowa State University

SUMMARY. We investigate the estimation of regression equations for a sample selected from a finite population. In all derivations, the finite population is treated as a sample from an infinite population. The regression coefficients are shown to be asymptotically normal, given mild assumptions. Relatively simple expressions for the covariance matrix of the regression coefficients are presented.

Procedures for estimating the structural parameters in the presence of response error are presented. Given knowledge of the response variance, the computations required to estimate the structural parameters and their standard errors are essentially equivalent to those required for the computation of the regression coefficient and its standard error in the absence of response error.

1. Introduction

In many scientific investigations, regression equations are computed for survey data. Examples of such studies include a consumption function for textiles where consumption is expressed as a function of income and occupation, a production function where farm output is expressed as a function of land and other inputs, and a function relating leisure activity to age and education.

The comparison of domain means is a kind of regression analysis wherein the independent variables take on only discrete values, such as 0 and 1. The survey literature, beginning in the 1950's (Yates, 1949) contains considerable material on the estimation of domain means and their differences. The terms, analytic studies (see Hartley, 1959) and analytic surveys, have become part of the survey statistician's vocabulary. Other than the work on analytic surveys there is very little literature dealing with the estimation of regression equations from survey data. One exception is the discussion of Deming (1950) wherein the problems of identifying the population for which inferences are desired is discussed.

Konijn (1962) considered the problem of estimating a regression equation from survey data. Given a population of N clusters of size M_i, $i = 1, 2, \ldots, N$, he assumed that the elements in each cluster were a random sample from an infinite population satisfying the usual linear model, $y = \alpha_i + \beta_i x$, $i = 1, 2, \ldots, N$.

*Journal Paper No. J-7975 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. Project No. 1806. This research was partly supported by a Joint Statistical Agreement with the U.S. Bureau of the Census, J.S.A. 74-1.

Editorial Note: Due to unavoidable circumstances 370, Part 3 could not be printed according to publication schedule. This unfortunate delay is regretted.

Konijn defined the parameters of interest to be the weighted average of the N cluster parameters, where the weights were the cluster sizes. Thus, the slope parameter of interest was $\beta = \left(\sum_{i=1}^{N} M_i \right)^{-1} \sum_{i=1}^{N} M_i \beta_i$. Given a sample of elements in each of a sample of n clusters, the suggested estimators were the weighted average of the n sample regression estimators computed for the n clusters. Under the model, the estimators are unbiased for α and β. The variance expression contained a component associated with the variability of the $\hat{\beta}_i$ as estimators of β_i and a component arising from the fact that a sample of the β_i is chosen from the finite population of β_i.

Frankel (1971) studied the empirical behavior of multiple regression coefficients computed from a cluster sample. The data were a sample of U.S. households collected by the U.S. Bureau of the Census in the March 1967 Current Population Survey. In Frankel's study, the objective of the regression analysis was the estimation of the finite population parameters as defined by the finite population moments. For the data studied by Frankel, the simple least squares estimators computed from an equal probability cluster sample displayed little bias, but the usual least squares estimate of variance underestimated the sampling variance by about 10 per cent.

In Frankel's study, variance estimators based on Taylor approximations, on balanced repeated replication and on jack-knife repeated replications, all gave reasonable estimates of the variance of the estimated regression coefficients.

There has been little explicit analytic treatment of the sampling properties of regression coefficients in the sampling literature. Frankel (1971) used the Taylor approximations to the variance suggested by Tepping (1968). In the general form presented by Tepping, these formulas are cumbersome for multiple regression equations. As we shall see, rather simple representations are possible for the estimated covariance matrix of the regression coefficients.

It is recognized that much data collected in sample surveys, particularly that collected from human respondents, are subject to measurement error. The U.S. Bureau of the Census (1972) has reported estimates of the response variance, as a percentage of total variance, that range from 0.5 to 40 per cent. Battese *et al.* (1972) report response variances of a similar magnitude for items associated with farm operations.

In a simple regression, the effect of uncorrelated response error in the independent variable is a reduction in the absolute value of the expected value of the regression coefficient relative to the expected value in the absence of response error. If the response errors in the independent and dependent

variables are correlated, the bias is increased or decreased, depending upon the signs of the error correlation and of the regression coefficient. Cochran (1968) and Chai (1971) discussed the effect of response variance on regression statistics. Fuller (1971) has investigated the properties of errors in variables estimators of regression parameters under the assumption of an infinite model with normal errors.

We shall consider the estimation of regression equations from samples selected from finite populations. It shall be assumed that the finite population is a random sample from an infinite population.

Assuming that the covariance matrix of the response errors is known, we present an estimator of the regression coefficients under a model that does not assume identically distributed response errors for each respondent.

2. Limiting distribution of the vector of regression coefficients

To investigate the behaviour of the estimator of the finite population regression coefficient, it seems necessary to make distributional assumptions or (and) to use large sample approximations. In this section, we investigate the limiting behavior of the estimated coefficients as both the sample size and the population size become large.

In investigating limiting properties of estimators, one must specify a sequence of finite populations and samples from these populations. One procedure, and that followed by Hájek (1960) and Madow (1948), is to treat the sequence of finite populations as a sequence of fixed numbers possessing certain well defined limiting properties. The required limits are roughly analogous to the existence of moments. An alternative approach, and the one we adopt, is to assume that the finite population is a random sample from a multivariate infinite population with finite fourth moments. It is also assumed that the covariance matrix of this multivariate population is positive definite.

We define the finite population vector of $(p+1)$ regression coefficients by

$$B = Q_N^{-1} H_N \qquad \ldots \quad (1)$$

and the infinite population vector of coefficients by

$$\beta = Q^{-1} H, \qquad \ldots \quad (2)$$

where the rs-th, $r, s = 0, 1, 2, \ldots, p$, elements of Q_N and Q are

$$q_{Nrs} = N^{-1} \sum_{i=1}^{N} x_{ir} x_{is},$$

$$q_{rs} = E\{x_{ir} x_{is}\},$$

respectively, the s-th elements of H_N and H are

$$h_{Ns} = N^{-1}\sum_{i=1}^{N} x_{is}y_i,$$

$$h_s = E\{x_{is}y_i\},$$

respectively, and $x_{i0} \equiv 1$.

The sample estimator of β, based upon a simple random sample of size n, is given by

$$b = Q_n^{-1}H_n, \qquad \ldots \quad (3)$$

where the rs-th element of Q_n is

$$q_{nrs} = n^{-1}\sum_{i=1}^{n} x_{ir}x_{is}, \quad r, s = 0, 1, 2, \ldots, p,$$

and the s-th element of H_n is

$$h_{ns} = n^{-1}\sum_{i=1}^{n} x_{is}y_i, \qquad s = 0, 1, 2, \ldots, p.$$

Theorem 1: *Let $\{\xi_n : n = 1, 2, \ldots\}$ be a sequence of finite populations, where ξ_n is a random sample of size N_n, $N_n > N_{n-1}$, selected from a p-dimensional infinite population. Assume the infinite population possesses finite fourth moments and a positive definite covariance matrix. Let a simple random nonreplacement sample of size n be selected from the n-th finite population, $n = 1, 2, \ldots$. Define $f_n = n/N_n$ and let*

$$\lim_{n\to\infty} f_n = f, \quad 0 \leqslant f < 1.$$

Then

$$n^{\frac{1}{2}}(b-B) \xrightarrow{\mathcal{L}} N(0, (1-f)Q^{-1}GQ^{-1})$$

as $n \to \infty$, where b is defined in (3), B is defined in (1), the rs-th element of G is

$$G_{rs} = E\{x_{ir}x_{is}e_i^2\},$$

and the population error, e, is defined by $e_i = y_i - \sum_{r=0}^{p} \beta_r x_{ir}$.

Proof: We may write[1]

$$b-\beta = Q_n^{-1}R_n,$$

$$B-\beta = Q_N^{-1}R_N.$$

[1] We simplify the notation in subsequent discussion by dropping the subscript n from N_n. For the same reason we have not subscripted b and B with n.

where $$R_n = \frac{1}{n}\left[\sum_{i=1}^{n} e_i, \sum_{i=1}^{n} x_{i1}e_i, \ldots, \sum_{i=1}^{n} x_{ip}e_i\right]'$$

and R_N is defined analogously.

Since the elements of Q_n are sample moments with variances of order n^{-1}, we have

$$Q_n - Q = O_p(n^{-\frac{1}{2}})$$

$$Q_N - Q = O_p(n^{-\frac{1}{2}})$$

and

$$n^{\frac{1}{2}}(b-B) = n^{\frac{1}{2}}Q^{-1}(R_n - R_N) + O_p(n^{-\frac{1}{2}})$$

$$= Q^{-1}\begin{bmatrix} n^{-\frac{1}{2}}(1-f_n)\sum_{i=1}^{n} e_i - f_n^{\frac{1}{2}}(1-f_n)^{\frac{1}{2}}(N-n)^{-\frac{1}{2}}\sum_{i=n+1}^{N} e_i \\ \vdots \\ n^{-\frac{1}{2}}(1-f_n)\sum_{i=1}^{n} x_{ip}e_i - f_n^{\frac{1}{2}}(1-f_n)^{\frac{1}{2}}(N-n)^{-\frac{1}{2}}\sum_{i=n+1}^{N} x_{ip}e_i \end{bmatrix} + O_p(n^{-\frac{1}{2}}).$$

Now, $E\{x_{ij}e_i\} = 0$, $j = 0, 1, 2, \ldots, p$, $E\{(x_{ij}e_i)^2\}$ is finite for $j = 0, 1, \ldots, p$, and the vectors, $(e_i, x_{i1}e_i, x_{i2}e_i, \ldots, x_{ip}e_i)$, $i = 1, 2, \ldots$ are independently and identically distributed. Therefore, letting $\lambda = (\lambda_0, \lambda_1, \lambda_2, \ldots, \lambda_p)$ be a vector of arbitrary real numbers (not all zero) the linear combination

$$S_{1n} = \sum_{j=0}^{p} \lambda_j n^{-\frac{1}{2}}(1-f_n)\sum_{i=1}^{n} x_{ij}e_i$$

converges in distribution to a normal random variable with mean zero and variance $(1-f)^2\lambda'G\lambda$ by the Lindeberg Central Limit Theorem. In a similar manner

$$S_{2n} = \sum_{j=0}^{p} \lambda_j f_n^{\frac{1}{2}}(1-f_n)^{\frac{1}{2}}(N-n)^{-\frac{1}{2}}\sum_{i=n+1}^{N} x_{ij}e_i$$

converges in distribution to a normal random variable with mean zero and variance

$$f(1-f)\lambda'G\lambda.$$

As S_{1n} and S_{2n} are independent, the result follows. □

It follows from Theorem 1 that $n^{\frac{1}{2}}(b-\beta) \xrightarrow{\mathcal{L}} N(0, Q^{-1}GQ^{-1})$. Thus, in analogy to analytic surveys (see Cochran, 1963, p. 37), one would not use the finite population correction if one were estimating the infinite population parameter.

It is of considerable interest that a consistent estimator of the variance of $n^{\frac{1}{2}}(b-\beta)$ can be constructed by estimating the matrix G by

$$\hat{G} = \frac{1}{n-p-1} \sum_{i=1}^{n} \hat{d}_{i\cdot} \hat{d}_{i\cdot}' \qquad \ldots (4)$$

where

$$\hat{d}_{i\cdot} = x_{i\cdot}' \hat{e}_i$$

$$\hat{e}_i = y_i - x_{i\cdot} b$$

$x_{i\cdot} = (x_{i0}, x_{i1}, \ldots, x_{ip})$ is the i-th row of the matrix used in constructing Q_n.

Theorem 2: *Let the sequence of samples and finite populations satisfy the assumptions of Theorem* 1 *and let* $\hat{G}$ *be defined in* (4). *Then plim* $\hat{G} = G$.

Proof: We have

$$\frac{1}{n} \sum_{i=1}^{n} \hat{d}_{ir} \hat{d}_{is} = \frac{1}{n} \sum_{i=1}^{n} x_{ir}[e_i - x_{i\cdot}(b-\beta)]\, x_{is}[e_i - x_{i\cdot}(b-\beta)]$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_{ir}\, x_{is}\, e_i^2 - \frac{1}{n} \sum_{i=1}^{n} e_i\, x_{ir}\, x_{is}\, x_{i\cdot}(b-\beta)$$

$$- \frac{1}{n} \sum_{i=1}^{n} e_i\, x_{is}\, x_{ir}\, x_{i\cdot}(b-\beta) + (b-\beta)' T (b-\beta),$$

where the sm-th element of T is given by

$$T_{sm} = \frac{1}{n} \sum_{i=1}^{n} x_{is}\, x_{it}\, x_{ir}\, x_{im}.$$

By Khinchine's weak law of large numbers (Rao, 1965 p. 92)

$$\frac{1}{n} \sum_{i=1}^{n} x_{is}\, x_{it}\, x_{ir}\, x_{im} \xrightarrow{P} E\{x_{is}\, x_{it}\, x_{ir}\, x_{im}\}$$

$$\frac{1}{n} \sum_{i=1}^{n} x_{ir}\, x_{is}\, e_i^2 \xrightarrow{P} E\{x_{ir}\, x_{is}\, e_i^2\} = G_{rs}.$$

As $b-\beta = O_p(n^{-\frac{1}{2}})$ the result follows. □

Although Theorem 2 presents a method of obtaining a consistent estimator of the variance in the absence of the usual assumptions of the linear model, it is clear that this estimator will be less efficient than the linear model estimator if $E\{e_i \mid x_i\} = 0$ and $E\{e_i^2 \mid x_i\} = \sigma^2$ for all x_i in the in finite population.

By use of the Central Limit Theorem of Appendix A, Theorem 1 can be extended to the regression coefficient estimated from two-stage stratified samples.

For a two-stage stratified sample selected from L strata, the rs-th element of the matrix Q_n is given by

$$q_{nrs} = \sum_{j=1}^{L} \frac{W_j}{n_j} \sum_{i=1}^{n_j} \frac{M_{ji}}{m_{ji}} \sum_{t=1}^{m_{ji}} x_{jitr}\, x_{jits}. \qquad \ldots (5)$$

where

W_j = fraction of population in stratum j,

n_j = number of primaries selected from stratum j,

M_{ji} = number of elements in primary i of stratum j,

m_{ji} = number of sample elements in primary i of stratum j,

x_{jitr} = the jit-th observation on the r-th independent variable.

Similarly, the s-th element of H_n is

$$h_{ns} = \sum_{j=1}^{L} \frac{W_j}{n_j} \sum_{i=1}^{n_j} \frac{M_{ji}}{m_{ji}} \sum_{t=1}^{m_{ji}} x_{jits}\, y_{jit} \qquad \ldots (6)$$

The elements of the matrix $n^{-1}\hat{G}$ are given by the usual formulas for the estimated covariances of means per primary for stratified two-stage sampling, using as the variables in the variance formulas

$$\hat{d}_{jitr} = x_{jitr}\,\hat{e}_{jit}. \qquad r = 1, 2, \ldots, p.$$

where

$$\hat{e}_{jit} = y_{jit} - \sum_{r=0}^{p} b_r\, x_{jitr}.$$

If the finite correction terms are ignored, the rs-th element of $\hat{G}$ is given by

$$n \sum_{j=1}^{L} \frac{W_j^2}{n_j} \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\bar{d}_{ji\cdot r} - \bar{d}_{j\cdot\cdot r})(\bar{d}_{ji\cdot s} - \bar{d}_{j\cdot\cdot s}). \qquad \ldots (7)$$

where

$$\bar{d}_{ji\cdot r} = \frac{M_{ji}}{m_{ji}} \sum_{t=1}^{m_{ji}} \hat{d}_{jitr}$$

$$\bar{d}_{j\cdot\cdot r} = \frac{1}{n_j} \sum_{i=1}^{n_j} \bar{d}_{ji\cdot r}$$

and

$$n = \sum_{j=1}^{L} n_j$$

3. Regression estimates in the presence of response error

We present a method for incorporating knowledge of the variance of the measurement error into the regression analysis.

We first generalize the model associated with equation (1) to include measurement error. We assume the vector $(y, x_1, x_2, \ldots, x_p)$ to be a random sample from a multivariate population with finite $4+\delta$, $\delta > 0$, moments. As before, we include the unit independent variable in the vector of x's if the intercept term appears in the model. Thus, for a model with intercept, the first x is always identically one. The infinite population regression parameters are defined by equation (2).

If a sample is selected from the population we do not observe y and x, but observe

$$Y_i = y_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

$$X_{ik} = x_{ik} + u_{ik}, \qquad k = 1, 2, \ldots, p,$$

where $w_i = (\varepsilon_i, u_{i1}, u_{i2}, \ldots, u_{ip})$ is the vector of response errors for the i-th observation. It is assumed that

$$E\{w_i \mid (y_i, x_{i1}, x_{i2}, \ldots, x_{ip})\} = 0$$

for all vectors $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$. It is also assumed that w_i is independent of w_j, $i \neq j$. We assume that the $4+\delta$, $\delta > 0$, moment of w_i is uniformly bounded. Note that we do not assume that the covariance matrix of w conditional on $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$ is constant for all $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$. The second moment matrix of $(Y, X_1, X_2, \ldots, X_p)$ is given by the sum

$$\begin{pmatrix} \sigma_y^2 & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} + \begin{pmatrix} \sigma_\varepsilon^2 & \Sigma_{\varepsilon u} \\ \Sigma_{u\varepsilon} & \Sigma_{uu} \end{pmatrix}$$

where the first matrix is the matrix of second moments of $(y, x_1, x_2, \ldots, x_p)$ and the second matrix is the covariance matrix of w. We assume that Σ_{xx} is nonsingular and that the covariance matrix of w is known.

Theorem 3: *Given the stated assumptions,*

$$n^{\frac{1}{2}}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \Sigma_{xx}^{-1} A \Sigma_{xx}^{-1})$$

as $n \to \infty$. *Further*

$$(\hat{\Sigma}_{XX} - \Sigma_{uu})^{-1} \hat{A} (\hat{\Sigma}_{XX} - \Sigma_{uu})^{-1}$$

is a consistent estimator of the variance of $\hat{\beta}$ *where*

$$\hat{\beta} = (\hat{\Sigma}_{xx}-\Sigma_{uu})^{-1}(\hat{\Sigma}_{xy}-\Sigma_{us})$$

$$\hat{\Sigma}_{xx} = \frac{1}{n}\sum_{i=1}^{n} X_{i.}'X_{i.}$$

$$\hat{\Sigma}_{xy} = \frac{1}{n}\sum_{i=1}^{n} X_{i.}'Y_{i.}$$

$$X_{i.} = (X_{i1}, X_{i2}, \ldots, X_{ip})$$

$$A = nE\{(\hat{\Sigma}_{xv}-\Sigma_{uv})(\hat{\Sigma}_{vx}-\Sigma_{vu})\}$$

$$\hat{\Sigma}_{xv} = \frac{1}{n}\sum_{i=1}^{n} X_{i.}'v_i$$

$$v_i = Y_i-X_{i.}\beta = e_i+\varepsilon_i-u_{i.}\beta$$

$$u_{i.} = (u_{i1}, u_{i2}, \ldots, u_{ip})$$

$$\Sigma_{uv} = \Sigma_{vu}' = E\{u_i'v_i\}$$

$$\hat{A} = \frac{1}{n-p}\sum_{i=1}^{n} \hat{d}_{i.}'\hat{d}_{i.}$$

$$\hat{d}_{i.} = X_{i.}'\hat{v}_i-\Sigma_{uv}$$

$$\hat{v}_i = Y_i-X_{i.}\hat{\beta}.$$

Proof: By the moment assumptions

$$\hat{\Sigma}_{xx} = \Sigma_{xx}+\Sigma_{uu}+O_p(n^{-\frac{1}{2}})$$

$$\hat{\Sigma}_{xy} = \Sigma_{xy}+\Sigma_{us}+O_p(n^{-\frac{1}{2}})$$

and, because Σ_{xx} is nonsingular

$$\hat{\beta}-\beta = O_p(n^{-\frac{1}{2}}).$$

Therefore

$$(\hat{\beta}-\beta) = \Sigma_{xx}^{-1}[\hat{\Sigma}_{xv}-\Sigma_{uv}]+O_p(n^{-1}).$$

Since $\hat{\Sigma}_{xv}-\Sigma_{uv} = n^{-1}\sum_{i=1}^{n}(X_{i.}'v_i-\Sigma_{uv})$ is a mean of independent random variables with finite $2+\frac{1}{2}\delta$ moments

$$n^{\frac{1}{2}}(\hat{\beta}-\beta) \xrightarrow{\mathcal{L}} N(0, \Sigma_{xx}^{-1}A\,\Sigma_{xx}^{-\frac{1}{2}})$$

2

by the Liapounov form of the multivariate central limit theorem. Now

$$\hat{v}_i = v_i - X_{i\cdot}(\hat{\beta}-\beta) = v_i + O_p(n^{-\frac{1}{2}})$$

and

$$X_{i\cdot}'\hat{v}_i\hat{v}_i X_{i\cdot} = X_{i\cdot}' v_i v_i X_{i\cdot} + O_p(n^{-\frac{1}{2}}).$$

The elements of the matrix

$$\frac{1}{n}\sum_{i=1}^{n}(X_{i\cdot}'v_i - \Sigma_{uv})(v_i X_{i\cdot} - \Sigma_{vu})$$

are the means of independent random variables with finite $1+\frac{1}{4}\delta$ moments, and it follows by Markov's weak law of large numbers (see Parzen, 1960 p.418) that

$$\frac{1}{n-p}\sum_{i=1}^{n}\hat{d}_{i\cdot}'\hat{d}_{i\cdot} \xrightarrow{P} A. \qquad \square$$

Corollary 3.1 : *Given a sequence of finite populations of size* $N = N_n = f^{-1}n$, *where* $0 < f < 1$, *selected as random samples from a population satisfying the assumptions of Theorem* 3, *and a sequence of samples of size* n *selected from these finite populations, then*

$$n^{\frac{1}{2}}(\hat{\beta}-B) \xrightarrow{\mathcal{L}} N(0, \Sigma_{xx}^{-1}(A-fG)\Sigma_{xx}^{-1}),$$

where A *is defined in Theorem* 3, G *in Theorem* 1 *and*

$$B = \left(\frac{1}{N}\sum_{i=1}^{N} x_{i\cdot}'x_{i\cdot}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_{i\cdot}'y_{i\cdot}\right)$$

as in Theorem 1.

Proof : Following the arguments of Theorems 1 and 3, we have

$$n^{\frac{1}{2}}(\hat{\beta}-\beta) = n^{\frac{1}{2}}\Sigma_{xx}^{-1}(\hat{\Sigma}_{xv}-\Sigma_{uv}-R_N)+O_p(n^{-\frac{1}{2}}),$$

where

$$R_N = \frac{1}{N}\sum_{i=1}^{N} x_{i\cdot}'e_{i\cdot}$$

Now

$$\operatorname{var}(R_N) - 2\operatorname{cov}(R_N, \hat{\Sigma}_{xv} - \Sigma_{uv}) = -\frac{1}{N}G.$$

Asymptotic normality follows by arguments analogous to those of Theorems 1 and 3. $\square$

If the covariance matrix of the errors is not known, but is estimated, the variance of the estimated coefficient vector is increased accordingly.

Corollary 3.2 : *Let*

$$\begin{pmatrix} s_z^2 & S_{zu} \\ S_{uz} & S_{uu} \end{pmatrix}$$

be an unbiased estimator of the measurement error covariance matrix distributed independently of $\hat{\Sigma}_{Xv}$. *Let*

$$d^{\frac{1}{2}} S_{uv} = d^{\frac{1}{2}}(S_{uz} - S_{uu}\beta) \xrightarrow{\mathcal{L}} N(0, C)$$

where $d = \eta^{-1} n$, η *a fixed positive number. Then, under the assumptions of Theorem* 3

$$n^{\frac{1}{2}}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \Sigma_{xx}^{-1}(A + \eta C))\Sigma_{xx}^{-1})$$

where

$$\hat{\beta} = (\hat{\Sigma}_{XX} - S_{uu})^{-1}(\hat{\Sigma}_{XY} - S_{uz}).$$

Proof : The result is immediate since

$$\hat{\beta} - \beta = \hat{\Sigma}_{xx}^{-1}[(\hat{\Sigma}_{Xv} - \Sigma_{uv}) - (S_{uv} - \Sigma_{uv})] + O_p(n^{-1})$$

and $\hat{\Sigma}_{Xv}$ and S_{uv} are independent by assumption. □

As in the case of regression estimation with no response errors, the procedures generalize to stratified multistage samples. If we assume that the response errors are independent between secondary units within the same primary unit as well as between secondary units in different primary units, the estimation formulas are immediate generalizations of (5), (6), and (7). The estimator is

$$\hat{\beta} = (\hat{\Sigma}_{XX} - \Sigma_{uu})^{-1}(\hat{\Sigma}_{XY} - \Sigma_{uz}),$$

where the rs-th element of the estimated second moment matrix of X,

$$\hat{\Sigma}_{XX,\,rs} = \sum_{j=1}^{L} \frac{W_j}{n_j} \sum_{i=1}^{n_j} \frac{M_{ji}}{m_{ji}} \sum_{t=1}^{m_{ji}} X_{jitr} X_{jits}$$

and the s-th element of the estimated cross moments of X and Y,

$$\hat{\Sigma}_{XY,\,s} = \sum_{j=1}^{L} \frac{W_j}{n_j} \sum_{i=1}^{n_j} \frac{M_{ji}}{m_{ji}} \sum_{t=1}^{m_{ji}} X_{jits} Y_{jit}.$$

The estimated variance of $\hat{\beta}$ is given by

$$\frac{1}{n}(\hat{\Sigma}_{XX}-\Sigma_{uu})^{-1}\hat{A}(\hat{\Sigma}_{XX}-\Sigma_{uu})^{-1},$$

where the rs-th element of $\hat{A}$ is

$$\hat{A}_{rs} = n\sum_{j=1}^{L}\frac{W_j^2}{n_j}\frac{1}{n_j-1}\sum_{i=1}^{n_j}(\tilde{d}_{ji\cdot r}-\bar{d}_{j\cdot\cdot r})(\tilde{d}_{ji\cdot s}-\bar{d}_{j\cdot\cdot s}) \qquad \ldots \quad (8)$$

and

$$\tilde{d}_{ji\cdot r} = \frac{M_{ji}}{m_{ji}}\sum_{t=1}^{m_{ji}}\hat{d}_{jitr}$$

$$\hat{d}_{jitr} = X_{jitr}\hat{v}_{jit}$$

$$\hat{v}_{jit} = Y_{jit}-\sum_{k=1}^{p}\hat{\beta}_k X_{jitk}$$

$$\bar{d}_{j\cdot\cdot r} = \frac{1}{n_j}\sum_{i=1}^{n_j}\tilde{d}_{ji\cdot r}.$$

If one assumes linearity, homogeneity of error variances, and normality, compact expressions for the covariance matrix of the estimates are possible (see Fuller, 1971).

To illustrate the computation of the errors in variables estimator, we use Frankel's (1971) data. A one-third systematic sample of the primaries of the original data was selected, and the log of the income of urban male heads of households aged 28-58 was regressed on age, age-squared, and years of education. This gives a sample of 952 primaries with nonzero entries in 30 strata for a total sample of 4020 elements. Estimates are given in Table 1. The sample was treated as a stratified cluster sample in the computation of standard errors. In computing the errors in variables estimates, it was assumed that the response variances for age, (age-43)^2 and education were 0.3, 91, and 3.00, respectively. It also was assumed that the response errors in the three variables were uncorrelated and that the response error in income was uncorrelated with that in age and education.

For age and age-squared, the response errors will be uncorrelated if the distribution of the response error in age is symmetric. The response error in age was estimated from data reported by Bailar (1968) and Palmer (1943). The response error in education was estimated from U.S. Bureau of the Census (1972 p. 49).

The estimated standard errors for the least squares regression were computed by using (7), and the estimated standard errors for the errors in variables regression were computed by using (8).

The adjustment of the covariance matrix for response errors resulted in a 10-per cent increase in the estimated coefficient for education, an amount approximately equal to three standard errors. The coefficient for age also increased considerably. These differences should be interpreted in light of the fact that age and education are among the items with the smallest response variances.

TABLE 1. LOG INCOME AS A FUNCTION OF AGE AND EDUCATION, URBAN MALE HEADS AGED 28-58

	age	$(\text{age}-43)^2$	education
least squares regression	.00213	−.00048	.0843
(standard errors)	.00116	.00013	.0032
errors-in-variables regression	.00236	−.00048	.0945
(standard errors)	.00117	00013	.0035

Hidiroglou (1974) has conducted a Monte Carlo study using the Frankel data. His results indicate that the large sample approximations were adequate for samples of 2 primaries per stratum for a population divided into 12 strata.

Acknowledgments

Michael Hidiroglou carried out the calculations for the example and read the manuscript. I wish to acknowledge the comments of T. M. F. Smith which led to improvements in the wording of the manuscript. J. J. Goebel and C. Asok read parts of the manuscript.

Appendix A

Central limit theorem for two-stage samples

In this Appendix we present a central limit theorem for the sample mean per primary unit of a two-stage sample from a finite population. The population mean per primary unit is defined to be the population total divided by the number of primary units in the population. Let $\{\xi_n : n = 1, 2, \ldots\}$ be a sequence of finite populations, where ξ_n contains N_n primary sampling units, $N_n > N_{n-1}$. The t-th observation in the s-th primary unit is denoted by

$$Y_{st} = u_s + e_{st}, \quad s = 1, 2, \ldots, N_n; \quad t = 1, 2, \ldots, M_s,$$

where M_s is the number of secondary units in the s-th primary, the random variable u_s is the 'primary component', and the random variable e_{st} is the 'secondary component'.

For the s-th primary it is assumed that the e_{st}, $t = 1, 2, \ldots, M_s$, are a random sample from an infinite population with zero mean, variance σ_{es}^2, and uniformly bounded $2+\delta$ moments $\delta > 0$. It is assumed that the vectors $(u_s, M_s, \sigma_{es}^2)$, $s = 1, 2, \ldots, N_n$, are a random sample from an infinite trivariate population with finite $4+2\delta$ moments. Let $\{g_{2s}\}$ be a fixed sequence such that $0 < g_{2s} \leqslant 1$, define m_s to be the smallest integer greater than or equal to $g_{2s}M_s$, and set

$$\Sigma_2 = \lim_{n\to\infty} \frac{1}{n} \sum_{s=1}^{n} E\left\{ \frac{M_s - m_s}{M_s m_s} \sigma_{es}^2 \right\}.$$

The primary sampling rate $f_1 = n/N_n$, $0 < f_1 < 1$, is assumed fixed, with the possible exception associated with the requirement that N_n be integer. Let the sample mean per primary be given by

$$\bar{y} = \frac{1}{n} \sum_{s=1}^{n} \frac{M_s}{m_s} \sum_{t=1}^{m_s} Y_{st}$$

and the population mean per primary by

$$\bar{Y} = \frac{1}{N} \sum_{s=1}^{N} \sum_{t=1}^{M_s} Y_{st},$$

where we have suppressed the subscript on N.

Theorem A : *Given the stated assumptions,*

$$n^{\frac{1}{2}}(\bar{y}-\bar{Y}) \xrightarrow{\mathcal{L}} N(0, \{1-f_1\}\Sigma_1+\Sigma_2),$$

where

$$\Sigma_1 = E\{M_s^2 u_s^2 + M_s\sigma_{es}^2\} - [E\{M_s u_s\}]^2.$$

Proof : We write

$$\begin{aligned}
\bar{d} &= \bar{y}-\bar{Y} \\
&= \frac{1}{n} \sum_{s=1}^{n} \frac{M_s}{m_s} \sum_{t=1}^{m_s} Y_{st} - \frac{1}{N} \sum_{s=1}^{N} \sum_{t=1}^{M_s} Y_{st} \\
&= \frac{1}{n} \sum_{s=1}^{n} M_s u_s + \frac{1}{n} \sum_{s=1}^{n} \frac{M_s}{m_s} \sum_{t=1}^{m_s} e_{st} - \frac{1}{N} \sum_{s=1}^{N} \sum_{t=1}^{M_s} (u_s + e_{st}) \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{s=1}^{n} M_s u_s + \sum_{s=1}^{n} \left(\frac{M_s}{nm_s} - \frac{1}{N}\right) \sum_{t=1}^{m_s} e_{st} \\
&\quad - \frac{1}{N} \sum_{s=1}^{n} \sum_{t=m_s+1}^{M_s} e_{st} - \frac{1}{N} \sum_{s=n+1}^{N} \sum_{t=1}^{M_s} (u_s + e_{st})
\end{aligned}$$

and

$$n^{\frac{1}{2}}\bar{Z} = n^{-\frac{1}{2}} \sum_{s=1}^{n} \{(1-f_1)\mu_s + (f_{2s}^{-1}-f_1) \sum_{t=1}^{m_s} e_{st} - f_1 \sum_{t=m_s+1}^{M_s} e_{st}\}$$

$$-f_1^{\frac{1}{2}}(1-f_1)^{\frac{1}{2}}(N-n)^{-\frac{1}{2}} \sum_{s=n+1}^{N} \left[\mu_s + \sum_{t=1}^{M_s} e_{st}\right],$$

where $f_{2s} = m_s/M_s$ and $\mu_s = M_s u_s - E\{M_s u_s\}$.

Now, the quantities, $\mu_s + \sum_{t=1}^{M_s} e_{st}$, $s = n+1, n+2, \ldots, N$, are independently distributed with variance $E\{(M_s u_s)^2 + M_s \sigma_{es}^2\} - [E\{M_s u_s\}]^2$ and bounded $2+\frac{1}{2}\delta$ moments. Therefore, by the Liapanouv Central Limit Theorem

$$\frac{\sum_{s=n+1}^{N} [\mu_s + \sum_{t=1}^{M_s} e_{st}]}{[(N-n)E\{\mu_s^2 + M_s \sigma_{es}^2\}]^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} N(0, 1).$$

Likewise,
$$(1-f_1)\mu_s + (f_{2s}^{-1}-f_1) \sum_{t=1}^{m_s} e_{st} - f_1 \sum_{t=m_s+1}^{M_s} e_{st}.$$

$s = 1, 2, \ldots, n$ are independently distributed with variance

$$(1-f_1)^2 E\{\mu_s^2 + M_s \sigma_{es}^2\} + E\{(f_{2s}^{-1}-1)M_s \sigma_{es}^2\}$$

and

$$\frac{\sum_{s=1}^{n} \left\{(1-f_1)\mu_s + (f_{2s}^{-1}-f_1) \sum_{t=1}^{m_s} e_{st} - f_1 \sum_{t=m_s+1}^{M_s} e_{st}\right\}}{\left[n(1-f_1)^2 \Sigma_1 + \sum_{s=1}^{n} E\{(f_{2s}^{-1}-1)M_s \sigma_{es}^2\}\right]^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} N(0, 1).$$

Since the two sums are independent for all n, the result follows. □

References

Bailar, B. A. (1968): Recent research in reinterview procedures. *J. Amer. Stat. Soc.*, 63, 41-63.

Battese, G. E., Fuller, W. A. and Hickman, R. D. (1972): Interviewer effects and response errors in a replicated survey of farm operators in selected Iowa counties. Report to Statistical Reporting Service, U.S. Dept. of Agri. Iowa State Univ., Ames, Iowa.

Chai, J. J. (1971): Correlated measurement errors and least squares estimation of the regression coefficient. *J. Amer. Stat. Assoc.* 66, 478-483.

Cochran, W. G. (1963): *Sampling Techniques.* Wiley, New York.

—— (1968): Errors of measurement in statistics. *Technometrics*, 10, 637-666.

Deming, W. E. (1950): *Some Theory of Sampling.* Wiley, New York.

FRANKEL, M. R. (1971): *Inference from Survey Samples.* University of Michigan, Ann Arbor.

FULLER, W. A. (1971): Properties of some estimators for the errors-in-variables model. Paper presented at the Econometric Society Meeting, New Orleans, December 1971.

HÁJEK, J. (1960): Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.*, **5**, 361-374.

HARTLEY, H. O. (1959): Analytic studies of survey data. Instituto di Statistica, Rome, Volume in onor e di Corrado Gini.

HIDIROGLOU, M. (1974): Estimation of regression parameters for finite populations. Ph.D. thesis. Iowa State Univ., Ames, Iowa.

KONIJN, H. S. (1962): Regression analysis in sample surveys. *J. Amer. Stat. Assoc.*, **57**, 590-606.

MADOW, W. G. (1948): On the limiting distribution of estimates based on samples from finite universes. *Ann. Math. Stat.*, **19**, 535-545.

PALMER, G. L. (1943): Factors in the variability of response in enumerative studies. *J. Amer. Stat. Assoc.*, **38**, 143-152.

PARZEN, E. (1960): *Modern Probability Theory and its Applications.* Wiley, New York.

RAO, C. R. (1965): *Linear Statistical Inference and Its Applications.* Wiley, New York.

TEPPING, B. J. (1968): Variance estimation in complex surveys. *Proceedings of the Social Statistics Section, Amer. Stat. Assoc.*, 11-18.

U.S. BUREAU OF THE CENSUS (1972): Evaluation and Research Program of the U.S. Censuses of Population and Housing 1960. Effects of different reinterview techniques on estimates of simple response variance. Series ER60, No. 11, Washington.

YATES, F. (1949): (3rd edition 1960) *Sampling Methods for Censuses and Surveys.* Griffin, London.

Paper received : June, 1974.
Revised : March, 1976.

72

Survey sampling errors with OSIRIS IV

Vinter, S., Ann Arbor, USA

Session C1/third paper
Sampling methods

Summary: the OSIRIS IV software system includes two programs for computing sampling error estimates derived from surveys with complex sample designs. The &PSALMS command produces sampling error estimates for ratios and ratio means, totals and differences of ratios. The &REPERR command produces sampling error estimates for means and regression statistics based on replication methods of variance estimation. Three alternative forms of replication methods are available: simple replication, balanced repeated replication, and jackknife repeated replication. The features of these two programs are described.

Keywords: Balanced half-sample, complex sample design, jackknife, primary selection, repeated replication, regression, sampling error, stratification, subclass, Taylor approximation, variance estimation.

INTRODUCTION

Complex sample designs, such as stratified multi-stage sampling, are widely used in survey research. The computation of sampling errors for survey estimates needs to take into account the type of sample design employed, and computer programs are required to enable the appropriate computations to be made. The OSIRIS IV software system, the current version of the OSIRIS system developed at the Institute for Social Research for the management and analysis of social science data, includes two programs for estimating sampling errors. This paper provides an introduction to these two programs, which are denoted by the &PSALMS and &REPERR commands in the OSIRIS IV system. The following sections describe the input data requirements, capabilities and limitations of these commands, and in doing so details their computational algorithms, special implementation features, and the output they produce.

Several techniques have been proposed for estimating sampling errors based on complex sample designs (see reviews by Kalton, 1977, and Shah, 1977). One approach approximates the variance of a statistic using the statistic's linear approximation, obtained by a Taylor's series expansion. This approach is most easily applied for relatively simple estimates and is adopted in the &PSALMS command to provide sampling error estimates for ratios and ratios means, differences between such ratios, and totals.

An alternative approach employs some form of the replicated variance estimation procedure. It is usually more costly to implement because of the

COMPSTAT 1980 © Physica-Verlag, Vienna for IASC (International Association for Statistical Computing), 1980

expense of computing a statistic for each replication. However, the attraction of this approach is the generality of its application; it is easily applied to a wide range of statistics due to the ease of computing the variance of a statistic from replications, irrespective of the complexity of the statistic. The &REPERR command uses the replicated variance estimation procedure to provide sampling error estimates for statistics produced in multiple regression analysis. It accomodates three alternative forms of replicated variance estimation: simple replication, balanced repeated replication (BRR), and jackknife repeated replication (JRR).

Kish and Frankel (1974) and Woodruff and Causey (1976) have examined and compared the approximate variance estimates produced by the Taylor expansion method, BRR, and JRR. In general, all three methods were found to provide good variance estimates.

For computational ease and generality, both commands assume that primary selections are sampled with replacement. In practice, however, they are usually sampled without replacement. The false use of the 'with replacement' assumption leads to an overestimation of the variance. But providing the first stage sampling fraction is small, the overestimation is slight. The 'with replacement' assumption yields a considerable savings in computing costs because the calculation of variance components from the subsequent stages of the design is unnecessary.

INPUT DATA REQUIREMENTS

The &PSALMS and &REPERR commands process OSIRIS datasets sequentially and use standard OSIRIS features for data retrieval, recoding, and output. A dataset consists of a dictionary file and data file. A dictionary file contains records describing variable attributes, including column location, storage type, and missing data codes. A data file contains one record per case with each record comprised of variables in fixed column locations. Most OSIRIS datasets are termed rectangular, being two dimensional with columns defining variables and each row representing a case. The commands can process OSIRIS hierarchical datasets, with data stored in a more complex structure, by reconfiguring the data into a rectangular form during data retrieval.

Required input specifications for both commands include the definition of a stratification variable and a sampling error computing unit (SECU) variable. These variables define the structure of the sample design to the commands.

The stratification variable divides the data into nonoverlapping strata that generally correspond to the stratification employed in the sample design.

Sampling error estimation requires at least two primary selections within each stratum. When the sample design does not provide this, the collapsed strata technique (Cochran, 1977) needs to be applied before the stratification variable is defined.

The definition of the SECU variable varies with different sample designs and variance estimation methods. It divides each stratum into unique, nonoverlapping units, and its formation is fundamental to the estimation procedure. With multi-stage sampling, each primary selection could constitute a separate SECU (with single-stage element sampling, each element can be a SECU). However, if the primary selections are numerous and small, it may be advantageous to join several selections to form a SECU. Primary selections may be combined within a stratum when a stratum contains many selections; alternatively, they may be combined across strata using the 'thickening zone' technique (Deming, 1960).

Different considerations apply between the two commands in determining appropriate definitions for the SECUs. The larger the number of SECUs defined, the greater the precision of the resulting variance estimators. Since the amount of computing in &PSALMS is relatively independent of the number of SECUs, generally as many SECUs as possible should be formed (i. e. each primary selection should be made into a separate SECU). However, the amount of computations in &REPERR is dependent on the number of replications formed, and replication formation depends to a certain extent on the number of SECUs defined, particularly with the BRR method. Therefore, it is sometimes desirable to employ a smaller number of SECUs for the &REPERR command.

An important feature in both commands is the capacity to account for empty SECUs, SECUs that correspond to one or more primary selections in the sample design but contain no valid cases (due to nonresponse or because the SECU contains no cases for a subclass during subclass analysis). In order that empty SECUs are not missed, all SECUs must be numbered consectively within each stratum, and the commands' setups must indicate the number of SECUs in each stratum.

Several features inherent to the OSIRIS IV system increase the flexibility of the commands. The powerful recoding facility, a large set of functions for creating and temporarily modifying variables, interfaces directly with both commands. The stratification and SECU variables may be created with this facility. Both commands permit data weighting, allowing each case to be assigned a different weight. A filtering capability is incorporated

in the commands to subset cases efficiently. Since cases are processed sequentially, they must be sorted by the stratification and SECU variables. The data can be presorted by the &COPYSORT command, or the sort option can be selected in the commands to indicate the data are to be sorted before analysis

THE &PSALMS COMMAND

&PSALMS uses the Taylor expansion method to estimate variances for ratio and ratio means, totals, and differences of ratios. Formulas for each statistic and its variance are found in Kish (1965). &PSALMS computes sums, sums of squares, and sums of cross products (SSQCP) for every variable in a ratio separately for each SECU. Missing data are deleted separately for each statistic under study. The variance and covariance contributions for every strata are computed by accumulating SECU SSQCP within strata according to one of three models: paired selection, successive difference, and multiple selection. Since stratum contributions are computed separately, different models may be employed for different strata in a single run.

The paired selection model is used for strata containing two SECUs. The successive difference model is available for cases where implicit stratification is obtained by systematic sampling from an ordered list (Kish, 1965); generally each selection is designated as a separate SECU, and all selections together form a single stratum. If systematic selection is used more than once for different parts of the sample, each set of selections defines a stratum. Since the order of the SECUs in this model reflects the implicit stratification and hence, affects the stratum variance computations, care must be taken to define SECUs in the order of selections employed in the systematic selection procedure. The multiple selection model, like the successive difference model, permits any number of SECUs per stratum. This model is available as a general model when strata contains a varying number of SECUs. Every strata must contain at least two SECUs in all models.

A missing SECU is entered as a zero in stratum contribution computations and does not increase the case count. Empty strata have no effect on overall variance estimation and are ignored. Printed diagnostics include SECU and strata counts and a count and listing of missing SECUs and strata.

A useful facility for subclass designation in &PSALMS enables a single run to yield sampling errors for a range of estimates for both the total sample and an unlimited number of subclasses.

The statistic, its standard error and variance, simple random sample standard error and variance, design effect, intraclass correlation, and case

and weight counts are included in the printout. The adequacy of the Taylor expansion variance estimate depends on a small coefficient of variation for the denominator for ratios as ratio means; the &PSALMS printout includes the coefficient of variation estimate and a warning when it is determined to be too high. &PSALMS permits the naming and numbering of each statistic under study for clear printed output, in addition to the inherent system features of automatic pagination and paging titling dating, and numbering. Additional printout options include variable sums and sums of squares for each SECU, a summary and extended table, and contributions to the variance for selected domains of analysis (sets of strata).

THE &REPERR COMMAND

&REPERR computes variances for means, correlations, regression coefficients, standardized regression coefficients, and multiple correlation coefficients using replication techniques. Replications, subset of SECUs that represent the overall population, are formed using a balanced replication procedure, jackknife procedure, or from user-specified lists of SECUs. Only one model may be used in a run.

Processing begins by combining sums, sums of squares, and sums of cross products of variables for each SECU to form replication totals in accordance with the selected model. Correlation matrices are computed and standard regression analyses are performed for the total sample and each replication. Estimates for each statistic from every replication and the total sample are combined to produce variance estimates (Frankel, 1971).

The BRR model requires that each stratum contains exactly two SECUs. One SECU is selected from each stratum to form a replication. This operation is repeated to form a set of replications that have the property of orthogonality (Kish and Frankel, 1970). This property is applied by using the procedure for producing the orthogonal matrices devised by Plackett and Burman (1946). The number of strata that can be accomodated by this procedure is limited to the range from 4 to 88. The complements of replications are also included as replications and used in variance computations, as discussed by Frankel (1974).

The jackknife model can be applied with any number of SECUs per stratum greater than one. A replication is formed by including all SECUs in the sample but one, and appropriately weighting the remaining SECUs in the stratum from which the one was deleted. This procedure can be repeated by deleting each SECU in turn, so that the total number of replications is equal to the number

of SECUs in the sample. If the costs of the analyses are too great using the total number of replications, a reduced set may be created by randomly selecting two SECUs in each stratum for deletion, thus forming two replications per stratum.

As an alternative to the BRR and JRR methods, the user may define replications by listing and weighting the SECUs to be included in each replication. This general model allows flexibility in replication formation and can be used for simple replicate sampling or methodological comparison of different replication procedures.

Missing strata are illegal and checked by the command. Missing SECUs are counted and documented, and processing continues with missing SECUs assumed empty and not contributing to replication totals.

Standard printout includes the statistic, its standard error, simple random sample standard error, design effect, and test statistic (corresponding to the t-statistic under assumptions of simple random sampling and normality). Print options available include: the listing of SECUs and weights by replication; SECU and replication univariate statistics, sums, sums of squares and cross products; and regression analysis by replication and for the total sample. A feature is available to create dummy variables from categorical measure for use as independent variables in regression analysis (see Draper and Smith, 1966, and Kish and Frankel, 1970)

REFERENCES

Cochran, W.G. (3rd ed., 1977). Sampling techniques. Wiley, New York.

Deming, W.E. (1960). Sample design in business research. Wiley, New York.

Draper, N.R., and Smith, H. (1966). Applied regression analysis. Wiley, New York.

Frankel, M.R. (1971). Inference from survey samples. Institute for Social Research, Ann Arbor.

Kalton, G. (1977). Practical methods for estimating survey sampling errors. Bull. Int. Statist. Inst., 43(3), 495-514.

Kish, L. and Frankel, M.R. (1970). Balanced repeated replications for standard errors. J.Amer.Statist.Assoc., 65, 1071-94.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples. J. R. Statist. Soc., B, 86, 1-37.

Plackett, R.L. and Burman, P.J. (1946). The design of optimum multi-factorial experiments. Biometrika, 33, 305-25.

Shah, B.V. (1977). Variance estimates for complex statistics from multi-stage sample surveys. In N.K. Namboodiri ed., Sampling survey and measurement. Academic Press, New York.

Woodruff, R. and Causey, B. D. (1976). Computerized method for approximating the variance of a complicated estimate. J. Amer. Statist. Assoc., 71. 315-321.

Appendix 2

Decription of OSIRIS program package

INSTITUTE FOR SOCIAL RESEARCH

Survey Research Center
Computer Support Group
P.O. Box 1248 48106

May 1980
Phone: (313) 764-4417

OSIRIS IV:

Statistical Analysis and Data Management Software System

SYSTEM OVERVIEW

OSIRIS IV is the current version of a software package which has been evolving for many years at the Institute for Social Research, University of Michigan. It makes use of the latest practical knowledge to reduce costs and provide increased capacity and flexibility in the areas of data management and statistical analysis. OSIRIS IV is designed to serve a broad community of users and has facilities for handling data collected for a wide range of purposes. In addition to the usual basic statistics and functions, such as cross-tabulations and classical regression and correlation analysis, several special techniques are available for handling nominal- and ordinal-scale data and for calculating sampling errors for complex designs. OSIRIS IV also has a full range of well integrated data management facilities; of special interest are the ability to handle weighted data, a powerful general purpose recording facility, matrix input and output, and hierarchical datasets with variable length records. Virtually any mode of data can be used directly in OSIRIS IV for up to 32,000 permanent variables. Among its other capabilities are facilities for:

- Online command documentation and indexing.
- Interactive setup interpretation
- Storing, retrieving, and modifying information about the structure of a dataset
- Displaying data
- Editing and correcting data
- Copying and subsetting data
- Transforming data values, through arithmetic and logical operations both within and across records
- Generating univariate and bivariate frequency distributions and related statistics
- Producing scatter plots
- Performing multiple regression analysis
- Performing univariate and multivariate analysis of variance
- Conducting analyses with multiple nominal- or ordinal-scale dependent and independent variables
- Searching among predictors for the greatest variance explanatory power (AID/SEARCH)
- Factor-analyzing data
- Performing cluster analysis.
- Multidimensional scaling.

Additional commands and documentation, developed within the Center for Political Studies to supplement and enhance the data structures capabilities of OSIRIS IV, include: extended capabilities for adding new and derived measures to a data structure, and procedures for subsetting the data structure; a capability for establishing what elements are present in the structure and in what proportion; and special documentation describing the generation, modification,

and use of hierarchical data structures.

Standard and consistent parameter keywords make OSIRIS IV easy to learn and use, and minimize the size and complexity of the required documentation. To use OSIRIS IV, the user supplies the following items, as appropriate:

- OSIRIS IV commands indicating which functions are desired and providing instructions to the OSIRIS IV monitor
- Data formatted either as an OSIRIS IV dataset or matrix
- Recode statements creating new variables or transforming existing ones
- Entry definitions indicating how groups of variables are to be assembled from a structured file
- Control statements specifying variables and parameters, optionally defining a subset of the data to be processed

Every variable in an OSIRIS IV dataset has a number and a fixed set of attributes associated with it: attributes such as the location of the variable within each record of the data file, the variable width, type, number of decimal places and the values to be treated as missing-data. This information is stored in a dictionary file, one record per variable. Once this information has been recorded in the dictionary, it need not be respecified by the user. Variables are then referenced in OSIRIS IV by their associated variable numbers. Dictionaries may easily be created or revised with the &DICT command..

Variables may be stored in a variety of modes:

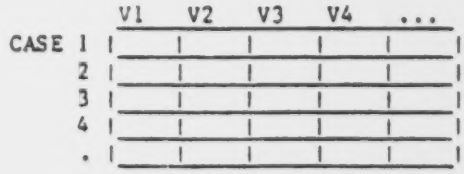
alphabetic
character numeric
floating-point binary
integer binary
packed decimal
zoned decimal

The storage mode of the variables need not be of concern except when first entering data into the OSIRIS IV system, as OSIRIS IV data management commands can handle data in any of the modes, and analysis commands can process all modes except alphabetic. OSIRIS IV commands which create new datasets will use the most efficient mode; however, &TRANS may be used to alter the storage mode of any or all nonalphabetic variables in a dataset.

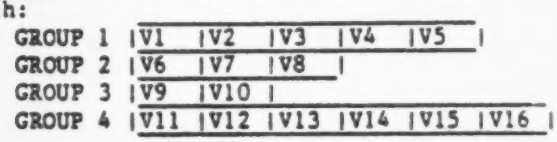
It should be noted that OSIRIS III datasets are compatible with OSIRIS IV; OSIRIS IV can both read and create datasets for use in OSIRIS III.

OSIRIS IV datasets have two possible configurations:

a. rectangular: all variables for one data case are stored in one record and each variable occupies the same relative location within each record:



b. structured: variables are collected into "groups" each with its own record length:



Selected variables from different groups may be joined together to create a rectangular record for a given run.

The first step in the creation of an OSIRIS IV rectangular dataset is to build an OSIRIS IV dictionary file, usually via the &DICT command. Once the dictionary has been created the data may be read directly by OSIRIS IV without any special "file building."

A structured dataset is built from individual rectangular files via the &SBUILD command. This type of dataset is used where there is not only a relationship between variables within a rectangular dataset, but also a relationship, usually hierarchical, between the various datasets. Each rectangular dataset becomes one or more groups in the structured dataset. A simple example is one rectangular dataset containing only household data and another containing additional data for individual members of each household merged by &SBUILD into a single structured dataset.

When a structured dataset is used in OSIRIS IV, the user gives instructions via the &ENTRY command as to how the groups are to be rearranged to create temporary rectangular records called "entries." This restructuring of the groups permits analysis to be performed on a wider range of entries than is possible with simple rectangular records and can thereby save numerous data management steps. The dictionary for the structured dataset may contain a default entry definition which is used to restructure the dataset when no other instructions are given via the &ENTRY command.

The advantage of a structured dataset is that for many data files, a more efficient storage mode is achieved in terms of space and cost of processing. This storage technique is more flexible and powerful than other data storage techniques, and permits larger datasets to be analyzed than in other systems.

OSIRIS IV is an open system; it is relatively easy in most instances to read data which are stored in character or binary form directly into an OSIRIS IV command. Another facet of the the system's openness is the ability to take any OSIRIS IV dataset and reformat it for use by other software. Thus, it is relatively easy to move outside the system; the data are not locked into OSIRIS IV. Finally, the user may add programs which use OSIRIS IV subroutines and hence use the common control statement language and OSIRIS IV datasets. Thus the software may be augmented to meet the user's special needs.

In addition to the basic user manual "OSIRIS IV: Statistical Analysis and Data Management Software System," there are many related publications of the Institute for Social Research that can serve as useful supplements. A list of these is attached.

HARDWARE REQUIREMENTS

The hardware requirements for OSIRIS IV are an IBM 360 or 370 computer, or an IBM-compatible machine such as an AMDAHL 470 V/6, with at least 150K bytes of

main storage, the equivalent of 1000 to 3000 tracks, 7294 characters each, of disk work space, and sufficient peripheral devices for user input and output files. The computer must be operated under MTS, the OS/360 or MVS operating system, or equivalent.

A periodic newsletter will be sent to all installations to communicate information about OSIRIS IV and its use. Finally, OSIRIS IV is not a static system; significant resources are being invested in its improvement and extension, and updates or new releases will be issued as changes are made. Commands currently being developed or planned include &AGGREG (aggregation), &CBKLIST (codebook listing), and &SPSSFILE (input SPSS files).

SYNOPSIS OF COMMANDS IN OSIRIS IV

Preparing Data for Input

&COPYSORT

Copies, reblocks and/or sorts OSIRIS IV and non-OSIRIS IV datasets.

&DICT

Creates, corrects, modifies, or adds to existing dictionaries, and adds code category labels to a dictionary.

&MATRIX

&MATRIX is used to enter one or more matrices into OSIRIS IV. Each matrix is assigned a unique number which is used to reference it in subsequent commands. Matrices which have been created by OSIRIS IV may simply be entered following an &MATRIX command. Matrices which have been created by other systems may also be read by providing the appropriate control statements.

Checking and Correcting Datasets

&CONCHECK

&CONCHECK used in conjunction with &RECODE provides a consistency check capability to test for illegal relationships between values for groups of variables. &CONCHECK takes user specifications indicating data inconsistencies from tests made in &RECODE and displays information allowing the user to locate each inconsistency. &RECODE and &TRANS or &FCOR can then be used to correct the inconsistencies.

&FCOR

&FCOR provides file correction capabilities for rectangular and structued OSIRIS datasets. It corrects values for any of the variables in any data case, adds a completely new record, or deletes an old one.

&MERCHECK

&MERCHECK detects and corrects merge errors for unit-record datasets (e.g., cards) such as missing decks, duplicate decks, or invalid datacards in datasets. The command produces a file in which each data case has the same

structure: a perfect merge of decks. The data for studies involving one deck of information per case should also be subjected to &MERCHECK because &MERCHECK will detect and correct multiple appearances of input data cards for any given case ID value, and will ensure that no cards foreign to the study have been included.

&WCC

&WCC verifies whether a set of variables has only legitimate data values and lists all invalid codes by case ID and variable number. Once the bad code values have been identified, they may be corrected with &FCOR.

Displaying Datasets

&DSLIST

&DSLIST is used to print a dictionary and/or a subset of variables from an OSIRIS dataset. &DSLIST is sometimes used to list temporary &RECODE result variables to check their correctness. A variety of formats is available.

Building and Modifying Structured Datasets

&ENTRY

With most structured files, the groups of variables created by &SBUILD can be combined in several ways. Each distinct combination of groups forms an entry, a single set of variables corresponding to a "case" in a rectangular file. &ENTRY allows the user to define or redefine the entry to be formed from the groups in the structured file, allowing the user to specify how the structured file is to be rectangularized.

&SBUILD

&SBUILD builds an OSIRIS IV structured dataset from one or more rectangular datasets. The basic unit of a structured dataset is a collection of related variables called a "group." A group has the same characteristics as a rectangular dataset: all the records are the same length and each variable is in the same relative location within each record. However, a structured dataset may contain many different groups, each with its own set of variables, and some logical relationship which ties them together.

&UPDATE

&UPDATE builds or updates OSIRIS IV rectangular or structured datasets from one or more OSIRIS IV rectangular or structured datasets. &UPDATE can add, delete, or replace cases or variables in a rectangular dataset, and add, delete, or replace groups, records, or variables in a structured dataset.

Transforming Datasets

&MATRANS

&MATRANS is used to change the type of a matrix, print a matrix, subset a matrix, and change variable numbers and names in a matrix.

&RECODE

A powerful recoding and variable transformation feature is available with almost all OSIRIS IV analysis and data management commands. The Recode facility can create new variables from any arithmetical combination of existing variables; can bracket or recode variables according to specified tables; and has several special features such as creating "dummy variables" and combination variables. In addition, a modest amount of aggregation and disaggregation may be accomplished via &RECODE.

&TRANS

&TRANS creates a rectangular OSIRIS dataset from specified input variables. &TRANS can convert the mode of the variables, and can also change the dictionary type for compatibility with other systems. It also allows the subsetting of cases. Additionally, &TRANS can be used to insert new variables created or modified by &RECODE into the dataset, thereby making permanent copies of them.

Frequency Distributions and Associated Statistical Measures

&SCAT

&SCAT is a bivariate analysis command which produces scatter diagrams, univariate statistics, and bivariate statistics. The scatter diagrams are plotted on a rectangular coordinate system; for each combination of coordinate values that appears in the data, the frequency of its occurrence is displayed. &SCAT is particularly useful for displaying bivariate relationships if the numbers of different values for each variable are large and the number of data cases containing any one value is small. If, however, a variable assumes relatively few different values in a large number of cases, &TABLES is more appropriate.

&TABLES

&TABLES produces univariate or bivariate frequency tabulations and percentages, and univariate statistics by stratum. (For univariate statistics, see also &USTATS.) &TABLES may also be used to produce quantiles and several nonparametric measures of association and significance for ordinal or nominal data. The Mann-Whitney U, the Kruskal-Wallis H, gamma, Kendall's tau a, b, c, lambda, lambda a, lambda b, Leik-Gove's D for nominal data (corrected), chi-square, Cramer's V, G-square, Gini coefficient and Lorenz plot, Goodman and Kruskal's tau a, b, and Cohen's Kappa.

&USTATS

&USTATS computes means, standard deviations, and minimum and maximum values for a given set of variables. Optionally, it will compute the same statistics for each variable for each specified subset.

Correlation and Regression Analysis

&MDC

&MDC computes Pearson product-moment correlation coefficients for all pairs of variables in a list, or for all combinations of variables, one of which is

from one list and another of which is from a second list or for selected pairs of variables. &MDC controls for input missing-data in one of two different ways: pair-wise or case-wise.

&PARTIALS

The n-th order partial-correlation coefficient (partial r) and the standardized partial regression coefficient (beta) are computed for each pair in a set of variables, holding all other variables constant. In addition, the multiple correlation coefficient (R) is computed for each variable using all the other variables as predictors. &PARTIALS is used in conjunction with &MATRIX or &MDC.

&REGRESSN

&REGRESSN will compute standard or step-wise multiple regressions with or without a constant term. It will accept interval or categorical (dummy) predictors. With the step-wise option, predictors may be forced into the regression before the step process begins. &REGRESSN will take as input an OSIRIS IV matrix or dataset. The latter may be weighted or unweighted and will be subject to a case-wise missing-data deletion. &REGRESSN may be used to "partial out" a subset of the predictors and print the remaining partial correlation matrix, prior to running the multiple regression with the full set of predictors. &REGRESSN will produce &RECODE control statements for computing residuals, if requested.

Analysis of Variance

&ANOVA

&ANOVA is a one-way analysis of variance command which performs an unlimited number of analyses using various independent and dependent variable pairs. &ANOVA will produce &RECODE control statements for computing residuals, if requested.

&MANOVA

&MANOVA performs univariate and multivariate analyses of variance and covariance, using a general linear hypothesis model. Up to twelve factors (independent variables) can be used. If more than one dependent variable is specified, both univariate and multivariate analyses are performed. &MANOVA performs an exact solution with either equal or unequal numbers of observations in the cells.

Multivariate Analysis Using Ordinal and Nominal Predictors

&DREG

&DREG provides a maximum likelihood regression capability for a dichotomous dependent variable using either a linear or logit model. &DREG may also be used to analyze multiway contingency tables whenever one dimension can be thought of as a dichotomous dependent variable.

&MCA

&MCA examines the relationships between several categorical independent vari-

ables and a single interval scaled dependent variable, and determines the effects of each predictor before and after adjustment for its intercorrelations with other predictors in the analysis. It also provides information about the bivariate and multivariate relationships between the predictors and the dependent variable. See Andrews, et al, Multiple Classification Analysis, for a complete description of the methodology used. &MCA will produce &RECODE control statements for computing residuals, if requested.

&MNA

&MNA performs a multivariate analysis of nominal-scale dependent variables. While the MCA technique described above assumes interval measurement of the dependent variable and an additive model, &MNA is designed to handle problems where the dependent variable is a nominal scale, the independent variables may be measured at any level, including nominal, and where any form or pattern of relationship may exist between any two variables. The program uses a series of parallel, dummy variable regressions derived from each of the dependent variable codes, dichotomized to a 0-1 variable.

&SEARCH

&SEARCH searches among a set of predictor variables for those predictors which most increase the researcher's ability to account for the variance or distribution of a dependent variable. The question, "what dichotomous split on which single predictor variable will give us a maximum improvement in our ability to predict values of the dependent variable?," embedded in an iterative scheme, is the basis for the algorithm used in this command. &SEARCH divides the sample, through a series of binary splits, into a mutually exclusive series of subgroups. Every observation is a member of exactly one of these subgroups. They are chosen so that, at each step in the procedure, the split into the two new subgroups accounts for more of the variance or distribution (reduces the predictive error more) than a split into any other pair of subgroups. The predictor variables may be ordinally or nominally scaled. The dependent variable may be continuous or categorical. &SEARCH is an elaboration of the OSIRIS III AID3 and THAID programs.

Factor Analysis and Multidimensional Scaling

&COMPARE

&COMPARE is based on Schonemann and Carroll's procedure for "fitting one matrix to another under choice of a central dilation and rigid motion." The technique rotates one configuration (the problem space) to the space of the other configuration (the target space) to achieve a least-squares fit. In seeking the best fit, the rotation is a "rigid motion," which maintains the orthogonality of the axes. A typical application is to compare the configurations produced by non-metric scaling analysis and factor analysis from the same data.

&FACTAN

&FACTAN provides a general factor analysis package that includes numerous options for the application of various factor analytic tools currently in use. Separate factor analyses may be performed on various subsets of variables in a single run.

&MINISSA

&MINISSA (Michigan Israel Netherlands Integrated Smallest Space Analysis) is a nonmetric multidimensional scaling command. The input to &MINISSA is a matrix of similarity or dissimilarity coefficients (e. g. , Pearson's r), the output is a geometric representation of the matrix in m dimensions. &MINISSA constructs a configuration of points in space using information about the order relations among the coefficients. Because it is usually possible to satisfy the order relations of the coefficients in fewer dimensions than would be necessary to reproduce the metric information, the technique is called smallest space analysis (SSA).

Cluster Analysis

&CLUSTER

&CLUSTER performs hierarchical cluster analysis. With input consisting of a symmetrical matrix of measures of similarities or dissimilarities, &CLUSTER successively partitions the dataset into a set of clusters as determined by a clustering criterion. Clustering methods include the minimum and maximum methods, the central vectors and coefficient alpha method for similarities, and the centroid distance and mean square error methods for dissimilarities.

Sampling Error Analysis

&PSALMS

Using the Taylor series approximation method, &PSALMS computes estimates and sampling errors for ratio means and totals for stratified clustered sample designs. &PSALMS accesses both weighted and unweighted data, and does not assume a simple random sample was taken. &PSALMS will optionally calculate sampling errors for parameters on subclasses of the dataset.

&REPERR

&REPERR computes estimates of regression statistics and their estimated sampling errors for data from clustered sample designs using repeated replication techniques. Replications are created using one of three methods: balanced half-sample, jackknife, or user-specified replications.

Supplemental Hierarchical Data Structure Support, developed by the Center for Political Studies Computer Support Group

&MERGE

&MERGE modifies an OSIRIS IV hierarchical dataset. &MERGE will correct variables, add new variables, add, delete, replace, or list occurrences (records), and selectively join two structured datasets together.

&SORTFLD

&SORTFLD provides specific information about the sort fields in an OSIRIS IV structured dataset. This information includes a description of the structure, a display of the sort fields, and an analysis of the occurrence of the logical pairs of groups in the data.

&STRANS

&STRANS will subset a structured dataset, keeping the structure intact. It is especially useful for creating a random subset of a very large dataset. &STRANS also permits new and derived measures to be added to a structure.

MAJOR DIFFERENCES BETWEEN OSIRIS IV AND OSIRIS III

1. Data Files

a) Hierarchical files with variable-length records may be created and used in OSIRIS IV. Such files can save space and execution time, and add flexibility for large-scale research data bases.

b) Virtually any kind of storage mode may be used for the data, including character, integer and floating-point binary, packed and zoned decimal.

c) Leading blanks and decimal points are permissible in the data.

2. Dictionary Files

Dictionary files are "type 5" in OSIRIS IV by default (required for hierarchical files); however, OSIRIS III dictionaries may be used without change in OSIRIS IV.

3. Codebook Records

a) "L" cards can be used to provide category labels for use in OSIRIS IV

b) OSIRIS III codebook records may be brought in and out of OSIRIS IV. However, OSIRIS III codebook records with text fields longer than 56 characters will cause additional records to be generated, in order to make room for the group number and expanded variable numbers required in OSIRIS IV. These additional records will be collapsed back into their original form if the dictionary is later converted back to OSIRIS III format with the &DICT command.

4. Matrix Files

OSIRIS IV matrix files such as created by &MDC and &FACTAN are automatically available to subsequent commands such as &REGRESSN AND &FACTAN.

5. RECODE

a) &RECODE statements must appear before the command which will use them.

b) Any mode recode may be used with any command. Decimal data will be rounded when integer mode RECODE is used.

c) Alphabetic recoding is possible.

6. "INTEGER" Programs

The distinction between "INTEGER" and "FLOATING-POINT" programs has been dropped. Commands for which only integer values are appropriate will automatically round decimal data to the nearest integer as needed. &RECODE could be used to scale such values to simulate OSIRIS III if desired.

7. Global Filters

-Numeric and alphabetic variables may be used.
-Leading zeroes do not have to be punched.

-Decimal points must be used as indicated by the dictionary.
-Filter statements can be of any character length.
-Parentheses may be used.
-The symbols > and < may be used.

8. Printout

-Page titles may be up to 100 characters long.
-Output is page-numbered and dated.
-Printout is more compact.

9. REPETITIONS

The use of REPETITION factors, which allow several analyses to be performed for different subsets of the data, has been greatly expanded and provides a facility analogous to "packets" in OSIRIS III.

NOTE: OSIRIS III remains available to users, as a software package with a variety of features not incorporated in OSIRIS IV. OSIRIS III is, however, a stabilized system with no further development currently underway.

DISTRIBUTION POLICY

OSIRIS IV is distributed on a not-for-profit basis by the Survey Research Center Computer Support Group (SRCCSG) at the Institute for Social Research, University of Michigan. A first-year fee and a subsequent yearly rental fee, specified in the order form, are charged to cover distribution, maintenance, and development costs. At the end of the first year, SRCCSG will automatically bill for the next 12 month period and continue to do so until a invoice is returned with a note asserting that OSIRIS IV is no longer in use.

I. Materials sent

Upon receipt of the order form and signed distribution agreement, complete with payment, SRCCSG will send the following materials to the requestor:

1. OSIRIS IV, including source modules, load modules which are intended to run under OS, MVS, or their equivalent on an IBM/360 or IBM/370, and installation implementation instructions, all on magnetic tape.

2. One (1) copy of the OSIRIS IV manual.

3. One (1) copy of the OSIRIS IV subroutine manual.

4. One (1) copy of OHDS: An Introduction to the OSIRIS Hierarchical Data Structures Capabilities.

II. Maintenance and Consultation

1. Enhancements, updates, and improvements to OSIRIS IV, including new releases, when and if developed for public distribution, will be sent automatically to all current OSIRIS IV installations.

2. Considerable effort has been made to make OSIRIS IV as trouble free to use and implement as possible, but should difficulties arise, a reasonable amount of free consultation by phone or letter will be available, with all phone charges to be paid by requestor.

3. Extended consultation on the use of OSIRIS IV may be available at $25.00/hour or by special arrangement with SRCCSG.

4. Maintenance and consultation will be provided only for the current release of OSIRIS IV, and may, but need not, be provided if the OSIRIS IV installation has modified or changed OSIRIS IV.

5. In the event of the loss or destruction of the installation's copy of the current OSIRIS IV, SRCCSG will replace it at a reasonable charge.

6. If serious errors are discovered, a revised tape will be sent out.

III. Distribution Agreement (repeated on back of order form)

1. The OSIRIS IV system is to be used by the requesting installation only on a single COMPUTER SYSTEM and for its own use, except as noted in section 3 below. The term single COMPUTER SYSTEM encompasses a multiprocessor system wherein the processors are located on the same site, a system wherein terminals are located off site, or other back-up system

located on the same site. Re-distribution of OSIRIS IV in whole or in part or any derivative thereof, externally or internally, to other computer systems or sites is prohibited.

2. No title or ownership rights to OSIRIS IV are transferred by this agreement.

3. If a commercial installation makes computer time which uses OSIRIS IV available to any other user, then the requesting installation agrees that it will pay SRCCSG, within thirty (30) days after the end of each calendar quarter, a 10% royalty on all charges to such other users made by the installation for machine related services for each computer job which utilizes OSIRIS IV.

4. All payments are exclusive of any tariffs, duties or taxes imposed or levied by any government or governmental agency. The requesting installation shall be liable for payment of all such taxes however designated, levied, or based on OSIRIS IV, its use, or on this agreement, including without limitation, state or local sales, use, and personal property taxes.

6. While OSIRIS IV has been carefully developed and tested for accuracy and proper functioning, SRCCSG, the Survey Research Center, the Institute for Social Research, or the University of Michigan cannot guarantee the accuracy or correctness of OSIRIS IV.

7. In no event shall the SRCCSG, the Survey Research Center, the Institute for Social Research, or the University of Michigan become liable to the requesting installation, or any other party, for any loss or damages, consequential or otherwise, including but not limited to time, money, or goodwill, arising from the use, operation or modification of OSIRIS IV by the requesting installation.

OSIRIS IV Distribution
Computer Support Group
Survey Research Center
Institute For Social Research
University of Michigan
Ann Arbor, Michigan 48106

OSIRIS IV ORDER FORM: Complete BOTH sides and return to above address.

Shipping Address:

Name _____

Firm/Institution _____

Street _____

City _____ State _____ Zip _____

Country _____ Telephone No. _____

This request must be accompanied by a check or purchase order. The current rates available are as follows:

first year	annual renewal	
$ 2400	$ 1800	basic fee
$ 1600	$ 1200	for government agencies and non-profit institutions
$ 1200	$ 900	for institutions granting academic degrees
$ 900	$ 675	for Inter-university Consortium for Political and Social Research (ICPSR) members

Overseas orders: add $25

|__| Check No._____ |__| Purchase Order No._____ Amount $_____

OSIRIS IV TAPE - indicate desired density:

9 track EBCDIC - odd parity tape written at

|__| 800 BPI |__| 1600 BPI |__| 6250 BPI

CHECK ONE CATEGORY:

|__| Degree Granting Institution

|__| Government or Non-profit

|__| Service Bureau

|__| Other:_____

HARDWARE:

Manufacturer_____

Model No._____

Memory size_____

Operating System_____

FOR SRCCSG USE ONLY:	Date Received	Version	Date Copied	Date Mailed
	_____	_____	_____	_____

DISTRIBUTION AGREEMENT

1. The OSIRIS IV system is to be used by the requesting installation only on a single COMPUTER SYSTEM and for its own use, except as noted in section 3 below. The term single COMPUTER SYSTEM encompasses a multi-processor system wherein the processors are located on the same site, a system wherein terminals are located off site, or other back-up system located on the same site. Re-distribution of OSIRIS IV in whole or in part or any derivative thereof, externally or internally, to other computer systems or sites is prohibited.

2. No title or ownership rights to OSIRIS IV are transferred by this agreement.

3. If a commercial installation makes computer time which uses OSIRIS IV available to any other user, then the requesting installation agrees that it will pay SRCCSG, within thirty (30) days after the end of each calendar quarter, a 10% royalty on all charges to such other users made by the installation for machine related services for each computer job which utilizes OSIRIS IV.

4. All payments are exclusive of any tariffs, duties or taxes imposed or levied by any government or governmental agency. The requesting installation shall be liable for payment of all such taxes however designated, levied, or based on OSIRIS IV, its use, or on this agreement, including without limitation, state or local sales, use, and personal property taxes.

6. While OSIRIS IV has been carefully developed and tested for accuracy and proper functioning, SRCCSG, the Survey Research Center, the Institute for Social Research, or the University of Michigan cannot guarantee the accuracy or correctness of OSIRIS IV.

7. In no event shall the SRCCSG, the Survey Research Center, the Institute for Social Research, or the University of Michigan become liable to the requesting installation, or any other party, for any loss or damages, consequential or otherwise, including but not limited to time, money, or goodwill, arising from the use, operation or modification of OSIRIS IV by the requesting installation.

The terms of this agreement are understood and accepted for the requesting installation by:

Name_____

Title_____

Signature_____

Date_____

OSIRIS IV AND RELATED DOCUMENTATION

Items listed below may be ordered using the attached order form.

OSIRIS IV: Statistical Analysis and Data Management Software System

A thorough description of each command and the overall system. The write-up for each command includes a general description, uses, functional relations to other commands, extended explanations of options and features, restrictions, input and output requirements, and sample setups. 1979. 250 pages, ring binder. Price: $15.00 ($9.00 for over-the-counter cash purchase from ISR supplies).

OSIRIS IV Subroutine Manual

The subroutines in the OSIRIS IV library are described in this manual. The functional characteristics are detailed as well as all entry points and calling sequences. This manual is useful to those wishing to modify existing OSIRIS IV commands or to add new commands. 1979. 160 pages, ring binder. Price: $13.00.

OSIRIS III, Vol. 5: Formulas and Statistical References by Laura Klem.

Although an OSIRIS III reference, this volume is reasonably applicable to the corresponding OSIRIS IV commands. An OSIRIS IV version is planned. 1974. 212 pages, loose-leaf, shrink wrapped. Price: $8.00.

Searching for Structure by John A. Sonquist, Elizabeth Lauh Baker, and James N. Morgan.

This monograph presents an approach to analysis of substantial bodies of microdata which is incorporated in the OSIRIS III program AID3. The OSIRIS IV &SEARCH command is a direct descendant of the AID3 program and several new features have been added. A new monograph is in progress, but is not expected to be available until late 1980. Revised edition, 1974. Price: $6.50 paperbound, $10.00 clothbound.

THAID: A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables, by James N. Morgan and Robert C. Messenger.

This monograph describes a technique for conducting multivariate analysis of categorical dependent variables. Although common in social research, such variables have, until recently, been difficult to handle with available statistical techniques. THAID describes a searching process which provides an efficient and effective means for sorting through a variety of analytic models to find the one most able to produce useful predictions. The technique calls for subgroups that differ maximally as to their distribution; it assumes neither additivity nor linearity, and so requires substantial samples of 1,000 or more cases. The OSIRIS IV command &SEARCH incorporates a modified version of this technique as an option. 1973. 98 pages. Price: $8.00 clothbound.

Multiple Classification Analysis: A Report on a Computer Program for Multiple Regression Using Categorical Predictors, by Frank M. Andrews, James N. Morgan, John A. Sonquist, and Laura Klem.

Multiple Classification Analysis is a technique for examining the interrelationships between several predictor variables and a dependent variable withn the context of an additive model. The OSIRIS IV &MCA command implements this tech-

nique, and is a direct descendent of the program described in this monograph. Revised edition, 1974. 105 pages. Price: $5.50 paperbound, $9.00 clothbound.

Multivariate Nominal Scale Analysis: a Report on a New Analysis Technique, by Frank M. Andrews and Robert C. Messenger.

This monograph describes a powerful additive technique for conducting multivariate analyses of categorical dependent variables. It is particularly useful for exploring the interrelationships among theoretical concepts tapped by one categorical dependent variable and substantial numbers of categorical independent variables. This technique, already successfully incorporated into the OSIRIS III program MNA, will be added soon to OSIRIS IV under the same name. 1973. 114 pages. Price: $5.00 paperbound, $8.00 clothbound.

Data Processing in the Social Sciences with OSIRIS, by Judith Rattenbury and Paula Pelletier.

This monograph is intended to guide researchers in the field of social science (or their assistants) through all the stages necessary for processing data with a computer. It introduces the basic components of computers and the different kinds of software necessary for using a computer and then discusses types of data and some of the preliminary data collection phases prior to computer processing. The monograph goes step by step through the data processing stages which must be accomplished before analysis can be undertaken. It outlines different kinds of analysis and describes the kinds of errors commonly made when using a computer for data processing, and gives some hints on how to avoid them. Although the examples used are designed for use with OSIRIS III, they may readily be translated for use with OSIRIS IV or another system. 1974. 245 pages. Price: $6.00 paperbound, $10.00 clothbound.

A Guide for Selecting Statistical Techniques for Analyzing Social Science Data, by Frank M. Andrews, Laura Klem, Terrence N. Davidson, Patrick M. O'Malley, and Willard L. Rodgers.

The Guide is intended to be useful to social scientists, data analysts, and graduate students who already have some knowledge of social science statistics. It presents a systematic but highly condensed overview of over 100 currently used statistics and statistical techniques and their uses. The core of the Guide--a decision tree--consists of 16 pages of sequential questions and answers which lead the user to the appropriate technique. 1974, third printing 1976. 38 pages. Price: $5.00 paperbound, five copies for $10.00 ($2.00 for over-the-counter cash purchase from ISR supplies).

The following documentation may be ordered from:

CENTER FOR POLITICAL STUDIES
Post Office Box 1248
Ann Arbor, Michigan 48106

OHDS: An Introduction to the OSIRIS Hierachical Data Structures Capabilities.

This manual offers a step-by-step presentation on the generation, modification, and use of hierachical data structures. It has many examples and diagrams, and is an important aid for understanding the new and powerful OSIRIS IV Hierarchical Data Structures. 1979. 100 pages. Price: $5.00.

Item 61 11:58 Apr09/81 31 lines
Neal Van Eck
New OSIRIS IV Manual, programs.

A new OSIRIS IV manual is now available from ISR Supplies. It reflects changes which have been made since the June 1980 update ti the sixth edition. The manual has been completely reviewed and changes have been made to clarify usage where needed. In particular, the &UPDATE write-up has been completely revised and includes more examples, per user suggestion. Substantial changes have also been made to the FUNDAMENTALS section, and a new section on structured files has been added to the DATA MANAGEMENT chapter. The &RECODE section has been completely restructured for ease of use. New programs which have been added are:

&AGGREG	Aggregates individual records across subsets defined by the user and computes summary statistics. (Available about May 1)
&CAP	Spatial configuration analysis.
&CBLIST	Prints OSIRIS dictionary-codebooks in a sophisticated format.
&FREE	Permits data to be read into OSIRIS IV in a format free manner.
&SASFILE	Reads a SAS internal file.
&SPSSFILE	Reads an SPSS internal file.

Other additions include an expanded description of the &SET command, a new storage type (half-byte integer binary--see &TRANS), a new option (one deck) for &MERCHECK, and a complete description of the "global delimiter"" option (in the FUNDAMENTALS section).